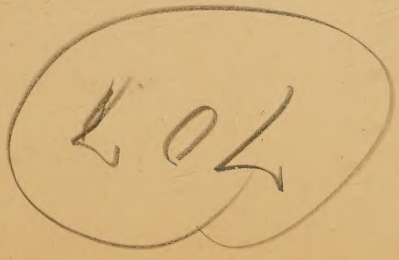
Ben-6-8333



AN INTRODUCTION
TO THE
MATHEMATICAL ANALYSIS
OF STATISTICS

AN INTRODUCTION
TO THE
MATHEMATICAL ANALYSIS OF STATISTICS

BY

C. H. FORSYTH

Assistant Professor of Mathematics
Dartmouth College

NEW YORK
JOHN WILEY & SONS, INC.
LONDON: CHAPMAN & HALL, LIMITED
1924

COPYRIGHT, 1924
BY
C. H. FORSYTH

PRESS OF
BRAUNWORTH & CO.
BOOK MANUFACTURERS
BROOKLYN, N. Y.

PREFACE

THIS book is offered as a textbook for a course in mathematics and not as a reference book for the statistician. Although a knowledge of the calculus is necessary for the proper appreciation of the various principles treated here, the applications are carefully restricted for the most part to those that do not require the calculus—except, possibly for a proper appreciation. All theory of a controversial character has been studiously omitted. It should therefore be possible for an instructor who is properly informed to employ certain parts of the book as a foundation for a course in statistics which presupposes only a most elementary knowledge of mathematics. The author set out originally to write a textbook on statistics in which the more advanced mathematical theory was to be relegated to an appendix but he gradually became convinced that a textbook worthy of use in a mathematical course should "go the whole way" and presuppose a knowledge of the calculus.

The author has departed considerably from the usual selection of topics—particularly in including numerical computation and finite differences. The only reason why numerical computation was included is that it was regarded as all important and yet is rarely taught or, at least, properly impressed upon the student. Courses in finite differences once enjoyed considerable popularity but have all but disappeared from the curricula of the colleges of the country. The theory is very valuable in statistical work and there seems to be no valid reason why it should be omitted here.

The main part of the book is given over for the most part to work connected with dispersion and this concept, once it is introduced, is continually emphasized throughout the rest of the book. There are many topics which it was a temptation to include but it was felt that the inclusion of some of these topics would affect seriously the continuity of the subject matter, which the author strived to maintain.

The author would indeed be lax in gratitude if he failed to give all credit for whatever of merit may be found herein to his former teachers Professor H. L. Rietz of the University of Iowa (formerly of the University of Illinois) and Professor J. W. Glover of the University of Michigan, to whom he is hopelessly indebted for both instruction and inspiration. He hastens, however, to assume for himself all responsibility for errors and blunders.

Last, but not least, thanks are due to the publishers and printers for their untiring efforts in the preparation of the book.

C. H. Forsyth.

CONTENTS

CHAPTER I

Errors and Numerical Computation . . . . . pages 1–11

CHAPTER II

Finite Differences 12–32

CHAPTER III

Interpolation 33–55

CHAPTER IV

Gamma and Beta Functions 56–63

CHAPTER V

Probability 64–78

CHAPTER VI

Averages and Aids in Their Computation 79–108

CHAPTER VII

Moments 109–135

CHAPTER VIII

The Normal Curve 136–166

CHAPTER IX

PAGES
THE BINOMIAL $(p+q)^n$. STATISTICAL SERIES 167–207

CHAPTER X

CORRELATION THEORY 208–232

INDEX 239–241

INTRODUCTION TO THE MATHEMATICAL ANALYSIS OF STATISTICS

CHAPTER I

ERRORS AND NUMERICAL COMPUTATION

1. Two Fundamental Facts.—Everyone who takes up the study of any form of applied mathematics must sooner or later come to appreciate at least two fundamental facts, which must always be kept in mind if anything like a satisfactory grasp of the subject is to be obtained. It is not the author's purpose to enter into a detailed discussion of these two facts, but simply to start the student pondering—if these ideas have never seriously occurred to him before—until he appreciates fully their reality and importance.

First of all, practically every application of mathematical theory, particularly to problems in natural and physical science, consists fundamentally in *idealizing* the given situation. That is, it is rarely, if ever, possible to find a mathematical expression or theory which explains or fits exactly the problem under discussion when all attending factors and influences are taken into consideration. Thus, the familiar mathematical function which is customarily associated with the law of falling bodies is based upon an idealization of the situation, and corrections must be applied to account for the effect of air currents, atmospheric pressure, etc. In the end the mathematical function plus all possible corrections can be expected to fit all experimental data only approximately.

Another fundamental fact which should be fully appreciated, and which is closely related to the fact mentioned above, is that all numerical measurements, or numerical results obtained by comparing the sizes of objects with a material unit used as a standard, are *relative* in value. It scarcely needs to be said that the absolute or exact length of a material object can never be determined by direct measurement, but there is frequent evidence that this fact is often forgotten or unappreciated. This fact applies not only to the measurement of length but also to such things as descriptions of locations and durations of time. The measurer is handicapped still further by the fact that the standard of measurement, chosen for its relative stability and rigidity, can never be perfect.

Man is little concerned with the impossibility of determining absolute measurements but is vitally concerned with relative measurements. No one, for example, spends much thought upon his own absolute location, but his location relative to food, drink, shelter, etc., probably concerns him more than anything else. Fortunately, the accuracy that man needs is also relative and so the one thing that is important in making numerical measurements is to see that the accuracy of the measurements shall be sufficient to meet the needs at the time.

2. Errors: Absolute and Relative Errors.—It is evident from what has been said that all measurements, as defined above, and all computations with measurements must involve errors of one kind or another. It follows that the absolute or exact values of these errors can never be determined; otherwise corrections could be made to give absolute measurements. Nevertheless, the *conception* of absolute error is useful, and so we shall define the *absolute error* of a measurement as the difference between the observed value and the absolute value. The absolute error is therefore positive or negative according as the observed value is too large or too small. We shall define, then, the *relative error* of a measurement as the ratio of the absolute error to the absolute value.

Although the exact value of the absolute error or of the

relative error of a measurement can never be known, it is a very important fact that their values can usually be *controlled*, in the sense that limiting values can be determined between which the true value must lie. For example, if an object is measured by a good foot rule and found to be 8 feet to the nearest foot, we know that the error can not be greater than 1 foot, and the size of the absolute error is thereby controlled. The relative error can not be greater than $\frac{1}{8}$ and its size is also evidently controlled.

3. Compensating and Accumulative Errors.—Errors are classified also with respect to their final and combined effect when their number in a given investigation is relatively large. Errors which tend to compensate or offset each other in the long run are called *compensating* errors. Large commercial concerns, which are accustomed to handling a large number of small sums of money daily, appreciate the fact that relatively trivial errors are usually compensating and tend in the long run to offset each other and leave only a small residuary discrepancy, which would scarcely warrant the expense of time and labor needed to investigate each trivial mistake. The term "errors" is often used to refer to much more than mere numerical mistakes. Thus, if we should tabulate as "errors" the deviations from the number 5 of the numbers found, say in the fifth decimal place of the values given in a table of logarithms, we should find that the excesses would tend to be offset in the long run by the deficiencies, or that the "errors" in this case would be compensating errors. It should be emphasized that effective compensation of such errors can be expected only when their number is relatively large.

Errors which tend in the long run to accumulate and give a relatively large combined error are called *accumulative* errors. Thus, if one should attempt to measure, by means of a defective foot rule, the distance between two points located a considerable number of feet apart, the errors would probably accumulate and give a relatively large error in the total distance found.

4. The Convention Followed in Expressing Measurements. —If we find by measurement that a length is 8.4 inches and so express it, we thereby imply that the measurement is to be regarded as correct to the nearest tenth of an inch. Such an implication accords with a generally adopted convention in regard to expressing the numerical results of a measurement, which specifies that *no more digits shall be written than are known to be correct, except whatever zeros may be needed to fill up the places of unknown digits immediately to the left of the decimal point or spaces immediately to the right of the decimal point when all the digits which are known to be correct are to the right of the decimal point.* Digits so written, which are known to be correct, are called *significant figures.* Thus, there are five significant figures in each of the numbers 302.02, 250.10, 0.0063284 and 2500.0, but only two in the number 93,000,000 because the zeros are used to fill up the places of digits which are unknown. When the distance from the earth to the sun is given as 93,000,000 miles the result is to be regarded as correct only to the nearest million of miles; that is the exact distance is to be regarded as lying between 92,500,000 miles and 93,500,000 miles. On the other hand, 2500.0 feet is to be regarded as correct to the nearest tenth of a foot and its exact value lies between 2499.95 feet and 2500.05 feet.

Sometimes we possess numbers expressing measurements which are given with greater accuracy than we care or need to use. Thus, suppose that we wish to express a measured length of 6.4 inches in terms of centimeters and we find in a table of equivalent lengths that 1 inch = 2.54001 centimeters. It would obviously be absurd to retain all of the significant figures in the latter number for the purposes of the problem just stated. When one or more digits of a number are dropped off, the number is said to be *rounded off* or simply *rounded.* A number is rounded off by dropping one or more digits at the right and, if the digit or digits dropped amount to more than one-half of one unit in the final place retained, by increasing the digit in that place by unity. Thus, the successive approxi-

mations to π obtained by rounding off 3.1416, one digit at a time, are 3.142, 3.14, 3.1 and 3.

The convention which has just been described refers particularly to numerical results of measurements. A slight modification of this convention is usually followed in expressing the numerical results of *computation* with measurements. This modification will be explained in the next section.

5. Computation with Rounded Numbers.—It has already been stated that although the exact value of an absolute error can never be determined it can be controlled, in the sense that certain limits can be placed upon the values that it may have, and that such limits can be indicated by following the generally accepted convention of expressing the numerical results of measurements which we have just explained. It is very important that the results of combining the numerical results of measurements by addition, subtraction, multiplication and division shall also be controlled, or that no illegitimate digits shall be retained as significant figures in such final results.

Two plans will now be offered for controlling the results of computation with measurements; one plan will be embodied in a set of general rules which may be easily followed and which will give results that are reasonably satisfactory in that they are *probably* correct; such rules can usually be employed also to show at once what computation may be avoided as useless. These rules will be stated in the next sections.

The other plan should be followed when a definite knowledge of the degree of the accuracy of the final result is essential, and calls for an analysis of each individual problem; it therefore involves a greater amount of labor and care. The plan consists simply in computing and comparing the maximum and minimum possible results, and can be explained best by a concrete illustration. As the maximum possible values of the measurements 39.2 inches and 18.3 inches are 39.25 inches (or so close to that number that no significant error will be committed by assuming it) and 18.35 inches the maximum possible product is then 39.25×18.35 or 720.2375. Similarly, the

minimum possible product is 39.15×18.25 or 714.4875, and it is evident on comparing these two results that if we should follow strictly the convention for expressing the numerical results of measurements our result when written would have only *one* significant figure and should be written 700. It is evident also, however, that if we kept only one significant figure considerable information would be ignored, for 717, the average of 720 and 714, differs by less than $3\frac{1}{2}$ units from either the maximum or the minimum possible results. Moreover, 717 is more probably correct than either of the extreme values 720 and 714, and is greatly preferable to 700 since the exact value can not be less than 714. It is customary, then, in expressing the numerical results of such a computation, to include also digits which are *probably* correct, especially when the possible deviation from such a result is small, say when the value of the possible deviation can not affect the digit preceding the last digit retained in the final result. Thus, there would be no justification for retaining four digits in the final result, say 717.4, since the possible deviation (2.8 or -2.9) would be too great.

Similarly, the maximum possible value of the quotient $18.3 \div 39.2$ would be $18.35 \div 39.15 = 0.4687$ and the minimum possible value of the quotient would be $18.25 \div 39.25 = 0.4650$. If the convention stated previously were strictly adhered to, it is evident that the quotient should be written 0.47. It is easily verified that the *probable* quotient is 0.467.

The procedure to be followed to determine the maximum and the minimum values of sums and differences is analogous and will suggest itself to the student.

It is rarely necessary to subject each computation to the individual analysis illustrated above, but it is very important that at least the general rules considered in the following sections should be kept in mind and followed.

EXERCISES

Show that

1. The sum of 13.26818, 138.36, 78.423, 7238.4289 and 6.324 could not be as large as 7474.82 or as small as 7474.79.

2. The difference between 362.34 and 47.26732 could not be as large as 314.09 or as small as 314.06.

3. The product of 34.68 and 4.6 to three digits could not be greater than 161 or less than 158.

4. The quotient of 36.4232 by 4.6 to two digits could not be greater than 8.0 or less than 7.8.

6. Addition and Subtraction.—The main end to be sought in numerical computation is that no more decimal places shall be retained in a final result than are correct or probably correct. Another goal which is not so important, but which every computer will naturally appreciate and seek, is the elimination of all computation which would probably have no effect upon the final result.

The following theorems are so self-evident that they scarcely call for any explanation: *the absolute error of the sum of two measurements is equal to the sum of the absolute errors of the measurements;* and *the absolute error of the difference between two measurements is equal to the difference between the absolute errors of the measurements.* Thus, if the absolute errors of the two measurements a and b are A and B respectively, then the sum of the two numbers is $(a+A)+(b+B)$ or $(a+b)+(A+B)$ and the error committed in taking $a+b$ as the sum is obviously $A+B$. Likewise, the error committed in taking $a-b$ as the difference between the two numbers is evidently $A-B$.

Attention is called to the fact that the two theorems stated above are practically equivalent; for, since the absolute error of a measurement may be either positive or negative, the sum (or difference) of the absolute errors of two measurements may prove to be a difference (or sum) of their absolute values. In any case we have the important corollary: *the absolute error of the sum or difference of two measurements whose absolute errors*

occur in different decimal places is equal approximately to the absolute error of the less accurate measurement. For, the absolute error of the more accurate measurement is relatively insignificant. We conclude then that *the sum or difference of two measurements should not be written to more decimal places than are given in the less accurate measurement.* To eliminate unnecessary computation it would seem desirable at first sight to round off the more accurate measurement only to within one place of the less accurate measurement before the addition or subtraction, but a little consideration will show that the same final result will be attained if the more accurate measurement is rounded off to the same decimal place as the less accurate measurement. As examples, the sum of 138.1 cms. and 26.032 cms. would be $138.1+26.0=164.1$ cms., and the difference between the same numbers would be $138.1-26.0=112.1$. It is easily verified that the sum could be as large as 164.18 or as small as 164.09, and that the difference could be as large as 112.118 or as small as 112.017.

If several measurements are to be added the measurements may be rounded off to within one decimal place of the least accurate measurements; *the sum should then be rounded off one more place.* The process is illustrated as follows:

136.421 cms.	136.4
28.3 "	28.3
321 "	321
68.243 "	68.2
17.482 "	17.5
	571.4 *Ans.* = 571 cms.

It is easily verified that the sum may be as large as 571.99 or as small as 570.90. It should be evident that the sum of a large number of measurements found in this way might differ considerably from the true sum if all or most of the absolute errors of the measurements should happen to be of the same sign. However, the larger the number of measurements the more likely it is, in the long run, that the errors will prove to be

compensating, and for that reason the sum so found is *probably* correct.

7. Multiplication and Division.—We shall now show that the rules to be followed in numerical computation with multiplication or division are based upon the idea of *relative* error.

If the relative errors of two measurements a and b are α and β, respectively, then a close approximation of the exact product of the two measurements is $(a+a\alpha)(b+b\beta)=ab+ab(\alpha+\beta+\alpha\beta)$, and the relative error committed in taking ab as the product is approximately $\alpha+\beta+\alpha\beta$. If, in addition, we ignore $\alpha\beta$ as relatively insignificant compared with either α or β we have the theorem: *the relative error of the product of two measurements is equal approximately to the sum of the relative errors of the measurements.*

Similarly, the relative error committed in taking a/b as the quotient of the two measurements is approximately $\alpha-\beta$; for, the absolute error committed is approximately

$$\frac{a+a\alpha}{b+b\beta}-\frac{a}{b}=\frac{a(\alpha-\beta)}{b(1+\beta)}.$$

But $\dfrac{\alpha-\beta}{1+\beta}$ differs very little from $\alpha-\beta$ and we have the theorem: *the relative error of the quotient of two measurements is equal approximately to the difference between the relative errors of the measurements.*

Just as the two theorems cited in connection with addition and subtraction were found to be practically equivalent, so the two theorems given above are practically equivalent and for the same reason, namely, that absolute errors, and consequently relative errors, may be positive or negative, and the sum (or difference) of two relative errors may prove to be a difference (or sum) of their absolute values. In any case, we have the important corollary: *the relative error of the product or quotient of two measurements having different numbers of significant figures is equal approximately to the relative error of the less accurate measurement.* It follows then that *such a product or quotient should not be written to more significant figures than*

appear in the less accurate measurement. Moreover, unnecessary computation will be avoided if the more accurate measurement is rounded off to within one of the number of significant figures contained in the less accurate measurement before multiplying or dividing. As examples, the product of 118.321 cms., and 12.1 cms., is $118.3 \times 12.1 = 1430$ sq. cms. The quotient of the two numbers would be $118.3 \div 12.1 = 9.78$. It is easily verified that the product could be as large as 1437 or as small as 1425, and that the quotient could be as large as 9.81 or as small as 9.74.

As the sums, differences, products and quotients obtained by the rules given in this and the preceding section usually prove to be close approximations of averages of the corresponding maximum and minimum values we shall refer to them as *probable values.*

EXERCISES

1. Find the maximum, minimum and probable values of the

(*a*) sum of 36.4823, 2.63, 783.4 and 36.488;
(*b*) difference between 38.426 and 22.1;
(*c*) product of 36.2 and 4.8;
(*d*) quotient of 3.64 by 4.6;
(*e*) quotient of 6.2 by 38.4.

2. Find the probable values of the

(*a*) sum of 26.834, 182.3, 5284.36 and 3.2648;
(*b*) difference between 324.86 and 189.7388;
(*c*) product of 836.4 and 0.06;
(*d*) product of 26.483 and 0.002;
(*e*) quotient of 26.483 by 0.002;
(*f*) quotient of 0.002 by 26.483.

3. The radius of a circle is found by measurement to be 34.6 inches. What is the circumference? ($\pi = 3.1415926536\ldots$)

4. Suppose that the radius of a circle were found by measurement to be 2.386274 inches. What is the circumference, correct to the nearest one-hundredth of an inch?

(Hint: A rapid computation with the first digits of π and of the radius will show that the result will have four significant figures.)

5. Express a measured length of 6.4 inches in terms of centimeters. (One inch = 2.54001 cms.)

6. Express a measured length of 34.2 cms. in terms of inches.

7. The area of a square field is 1286 square feet, correct to the last place. What is the length of each side, expressed to the maximum number of places?

CHAPTER II

FINITE DIFFERENCES

8. Discrete Values.—Everyone who has had any experience in plotting graphs is familiar with the fact that no matter how closely two or more points are plotted there are always other points which could be plotted between those already plotted—that is, if the function is continuous in that interval. In other words, there are no "vacant" spaces between two points on such a graph. There are many functions, however, which have little or no meaning except for particular values—such as integral values, of the independent variable. Thus, the number of people living at various ages in a given community is limited to integral values; fractional values, for example, would have no meaning. Likewise, n^2 regarded as the formula for the sum of any number of terms of the series $1+3+5+7+$ etc., has no meaning except for integral values of n. Values of a variable which are thus restricted are said to be *discrete;* that is, they are separated by "vacant" spaces.

Any set of numerical results of measurements made by comparison with a material standard, or computations with such measurements, must necessarily be discrete, because absolute values can never be determined. Tabulated values—logarithmic, trigonometric, financial, etc.—must therefore be discrete. In practically all of these cases the corresponding mathematical expression or function is either unknown or too complicated for purposes of valuation every time a particular value is desired; and therefore advanced mathematical principles are employed for computation, or the skill and experience of an expert are employed, to experiment and set up tables of these

values for ready reference. Since these values must be discrete it follows that not all of the values which will be needed can be expected to be included, especially if great accuracy is essential. The theory of finite differences is singularly appropriate for idealizing the law of uniformity of a set of discrete values, that is, for ascribing an artificial law of uniformity which can be employed readily to find other values not included in the table which will be sufficiently approximate to meet the needs.

9. Definition of a Finite Difference.—If, as is customary in the theory of finite differences, we denote a function of x by u_x (corresponding to $f(x)$ employed in ordinary mathematical analysis), the finite difference of u_x, denoted by the symbol Δu_x, may be defined by the *general* relation

$$\Delta u_x = u_{x+h} - u_x,$$

where h is any real constant.

It will be found possible, however, for all our immediate purposes, to adopt the increment h as the unit of measurement or, what amounts to the same thing, to assume h to be unity. We shall therefore define the finite difference of u_x by the particular relation

$$\Delta u_x = u_{x+1} - u_x. \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

We shall restrict all applications of the finite difference to applications of this particular relation. Thus, the finite difference of x^2 is

$$\Delta x^2 = (x+1)^2 - x^2 = 2x+1.$$

Similarly

$$\Delta a^x = a^{x+1} - a^x = a^x(a-1),$$

and

$$\Delta \log x = \log (x+1) - \log x = \log \frac{x+1}{x}.$$

Second, third and higher differences are merely successive

differences of the first and are designated as shown in the following examples. Thus,

$$\Delta x^3 = (x+1)^3 - x^3 = 3x^2 + 3x + 1,$$

$$\Delta^2 x^3 = \Delta(3x^2+3x+1) = 6x+6,$$

$$\Delta^3 x^3 = 6,$$

$\Delta^4 x^3$ and higher differences of $x^3 = 0$.

EXERCISES

Find the first, second and third differences of

1. x^4.
2. $x^3 - 4x^2 + 2x + 1$.
3. 2^x.

Show that

4. $\Delta C (C = \text{a constant}) = 0$.
5. $\Delta C u_x = C \Delta u_x$.
6. $\Delta C \cdot a^{mx+b} = C(a^m - 1)a^{mx+b}$.
7. $\Delta^3 \log x = \log \dfrac{(x+1)^3(x+3)}{x(x+2)^3}$.
8. $\Delta(u_x + v_x + w_x + \text{etc.}) = \Delta u_x + \Delta v_x + \Delta w_x + \text{ etc.}$
9. $\Delta u_x v_x = v_{x+1}\Delta u_x + u_x \Delta v_x$.
10. $\Delta \dfrac{u_x}{v_x} = \dfrac{v_x \Delta u_x - u_x \Delta v_x}{v_x v_{x+1}}$.
11. $\Delta^n x^n = n!$ where n is a positive integer and $n! = n(n-1) \ . \ . \ . \ 3 \cdot 2 \cdot 1$.
12. $\Delta^m x^n = 0$ where m is the greater of the two positive integers m and n.
13. Prove the identity: $u_{x+1}\Delta v_x + v_x \Delta u_x = v_{x+1}\Delta u_x + u_x \Delta v_x$.

10. Tabulations of Differences.—It is often desirable to employ some form of the following scheme of tabulating differences:

$$
\begin{array}{llll}
u_x & & & \\
 & \Delta u_x & & \\
u_{x+1} & & \Delta^2 u_x & \\
 & \Delta u_{x+1} & & \Delta^3 u_x \\
u_{x+2} & & \Delta^2 u_{x+1} & \\
 & \Delta u_{x+2} & & \\
u_{x+3} & & &
\end{array}
\Bigg\} \quad \ldots\ldots \quad (A)
$$

etc.,

where each expression is the difference between the two expressions immediately to the left—the lower minus the upper.

If zero is substituted for x in the scheme just given we have

$$
\begin{array}{llll}
u_0 & & & \\
 & \Delta u_0 & & \\
u_1 & & \Delta^2 u_0 & \\
 & \Delta u_1 & & \Delta^3 u_0 \\
u_2 & & \Delta^2 u_1 & \\
 & \Delta u_2 & & \\
u_3 & & &
\end{array}
\Bigg\} \quad \ldots\ldots \quad (B)
$$

etc.,

which is equivalent graphically to a translation of the y-axis in scheme (A) x units to the right. As this transformation is common in analytic work and often simplifies computation, scheme (B) can usually be made to fit any given situation and so will prove very useful in much that follows.

The terms u_0, Δu_0, $\Delta^2 u_0$, etc., are called the *leading term* and *leading differences* of u_x and form what is called the *principal diagonal.* As an illustration, if $u_x = x^3$,

$$
\begin{array}{llllll}
u_0 = 0 & & & & & \\
 & \Delta u_0 = 1 & & & & \\
u_1 = 1 & & \Delta^2 u_0 = 6 & & & \\
 & \Delta u_1 = 7 & & \Delta^3 u_0 = 6 & & \\
u_2 = 8 & & \Delta^2 u_1 = 12 & & \Delta^4 u_0 = 0 & \\
 & \Delta u_2 = 19 & & \Delta^3 u_1 = 6 & & \\
u_3 = 27 & & \Delta^2 u_2 = 18 & & & \\
 & \Delta u_3 = 37 & & & & \\
u_4 = 64 & & & & &
\end{array}
$$

etc.,

which is usually written in a more abbreviated form, such as

$u_x=$	x^3	Δu_x	$\Delta^2 u_x$	$\Delta^3 u_x$	$\Delta^4 u_x$
	0				
		1			
	1		6		
		7		6	
	8		12		0
		19		6	
	27		18		
		37			
	64				
	etc.				

The leading term and differences of x^3 are then 0, 1, 6, 6, 0, etc.

EXERCISES

Arrange in tabular form (B) the values and differences of the following functions:

1. $3x+2$.

2. $x^2-7x+12$.

3. x^3.

4. $x^4-3x^3+4x^2-6x+2$.

5. How does the tabular arrangement of the function 4^x differ fundamentally from those of the functions given in the preceding exercises?

6. Tabulate the values $\log 460=2.662758$, $\log 462=2.664642$, $\log 464=2.666518$ and $\log 466=2.668386$ and their differences. Would you say that the differences finally vanish *absolutely*? Explain.

7. Check the relations

$$(a)\quad u_{x+n}=u_x+n\Delta u_x+\frac{n(n-1)}{2}\Delta^2 u_x+\text{ etc.},$$

$$(b)\quad \Delta^n u_x=u_{x+n}-nu_{x+n-1}+\frac{n(n-1)}{2}u_{x+n-2}-\text{ etc.},$$

for $n=1$, 2 and 3 using the relations shown in table (A).

11. Rational Integral Functions.—A rational integral function is a function that can be written in the form

$$ax^n+bx^{n-1}+cx^{n-2}+ \text{ etc.},$$

where the coefficients are real and the exponents are positive integers. Thus, x^3+4x^2-6x+2 is a rational integral function.

Since each difference of x^n, where n is a positive integer, lowers the degree of that function by unity, and since

$$\Delta(u_x+v_x+w_x+\text{etc.})=\Delta u_x+\Delta v_x+\Delta w_x+ \text{ etc.},$$

and

$$\Delta(\text{constant})=0,$$

it follows that each difference of a rational integral function lowers the degree of the function by unity and that, after a while, differences are finally reached which are zero. This fact distinguishes rational integral functions from all other functions. It is left for the student to show that the third difference of x^3+4x^2-6x+2 is constant and that higher differences are therefore zero.

12. Factorials.—If we define the expressions $x^{(n)}$ and $x^{(-n)}$ by the relations

$$x^{(n)}=x(x-1)(x-2)\ldots(x-n+1),$$

$$x^{(-n)}=\frac{1}{x(x+1)(x+2)\ldots(x+n-1)},$$

it is easily verified that

$$\Delta x^{(n)}=nx^{(n-1)}, \quad \ldots\ldots\ldots\ldots \quad (2)$$

and

$$\Delta x^{(-n)}=-nx^{(-n-1)}. \quad \ldots\ldots\ldots\ldots \quad (3)$$

Thus,

$$\Delta x^{(3)}=\Delta x(x-1)(x-2)=3x^{(2)}=3x(x-1).$$

Expressions of the form $x^{(n)}$ and $x^{(-n)}$ are called *factorials* in the theory of finite differences and will be found to play a very important part in the work of the succeeding pages. Such factorials should not be confused with the factorial $n!=n(n-1)(n-2)\ldots 3\cdot 2\cdot 1$, common in ordinary mathematical analysis.

EXERCISES

1. Verify formula (2).

2. Verify formula (3).

3. Given ${}_xC_r$ (the number of different combinations of x things taken r at a time) $=\dfrac{x(x-1)(x-2)\ldots(x-r+1)}{r!}$, find the first, second and r-th differences and express the results in terms of the symbol ${}_xC_r$.

If we define $u_x^{(n)}$ and $u_x^{(-n)}$ by the relations

$$u_x^{(n)}=u_xu_{x-1}u_{x-2}\ldots u_{x-n+1},$$

$$u_x^{(-n)}=\frac{1}{u_xu_{x+1}u_{x+2}\ldots u_{x+n-1}},$$

show that

4. $\Delta u_x^{(n)}=(u_{x+1}-u_{x-n+1})u_x^{(n-1)}$.

5. $\Delta u_x^{(-n)}=(u_x-u_{x+n})u_x^{(-n-1)}$.

6. $\Delta(ax+b)^{(n)}=an(ax+b)^{(n-1)}$.

Check by using the result given in Exercise 4.

7. $\Delta(ax+b)^{(-n)}=-an(ax+b)^{(-n-1)}$.

Check by using the result given in Exercise 5.

8. Show that $(x+k)^{-(n)}=(x+k+n-1)^{(-n)}$,

and $(x+k)^{-(-n)}=(x+k+n-1)^{(n)}$.

9. Find $\Delta(x+k)^{-(n)}$ and $\Delta(x+k)^{-(-n)}$.

13. Newton's Formula.

Since $u_1 = u_0 + \Delta u_0,$

$$u_2 = u_1 + \Delta u_1 = (u_0 + \Delta u_0) + (\Delta u_0 + \Delta^2 u_0)$$

$$= u_0 + 2\Delta u_0 + \Delta^2 u_0,$$

$$u_3 = u_0 + 3\Delta u_0 + 3\Delta^2 u_0 + \Delta^3 u_0,$$

etc.,

we have, in general

$$u_x = u_0 + x\Delta u_0 + \frac{x(x-1)}{2}\Delta^2 u_0 + \frac{x(x-1)(x-2)}{3!}\Delta^3 u_0 + \text{etc.}, \quad . \quad (4)$$

where, it is to be noted, the coefficients follow the binomial law. Expansion (4) is called *Newton's formula.*

Newton's formula is so important that we shall derive it in another and more rigorous manner, as follows. Let us consider those functions which it is reasonable to assume can be expanded in the form

$$u_x = a + bx + cx^{(2)} + dx^{(3)} + \text{etc.} \quad . \quad . \quad . \quad . \quad (C)$$

Then, evidently

$$a = u_0.$$

Moreover

$$\Delta u_x = b + 2cx + 3dx^{(2)} + \text{etc.},$$

and

$$b = \Delta u_0.$$

Similarly,

$$c = \frac{\Delta^2 u_0}{2}, \quad d = \frac{\Delta^3 u_0}{3!}, \text{ etc.}$$

Substituting these expressions for the coefficients in equation (C), we obtain Newton's formula in the form

$$u_x = u_0 + x\Delta u_0 + x^{(2)}\frac{\Delta^2 u_0}{2} + x^{(3)}\frac{\Delta^3 u_0}{3!} + \text{etc.} \quad . \quad . \quad (4')$$

It is possible to expand most functions by Newton's formula, but rational integral functions are unique in that their expansions always involve only a finite number of terms. As an illustration, let us expand x^3. As the leading term and differences were found previously to be 0, 1, 6, 6, 0, etc., these values may be substituted in Newton's formula to give

$$x^3 = x + 3x^{(2)} + x^{(3)} = x + 3x(x-1) + x(x-1)(x-2).$$

It is easily verified that the expression on the right can be reduced to x^3.

The general term of a sequence (or of a series), such as

0, 1, 6, 6, 0, 0, etc.,

can be determined in like manner, *provided the general term is rational integral.* Thus, if the terms of the series

$$1+2+5+10+ \text{ etc.},$$

are differenced, the leading term and differences are found to be 1, 1, 2, 0, etc., and if these values are substituted in Newton's formula we obtain the general term $1+x+x^{(2)}$ or $1+x^2$. Can the general term of the series, $1+3+3^2+3^3+$ etc., be determined in this manner? Explain.

14. The General Term of a Sequence or of a Series.—In ordinary mathematical analysis we usually understand the general term of a sequence or of a series to be the expression which gives the value of that term when the number of that term is substituted in the expression. In the theory of finite differences it is usually more natural to assume that a series is of the form

$$u_0+u_1+u_2+u_3+ \text{ etc.}$$

or that the general term will give the first term for $x=0$, the

second term for $x=1$, etc. According to that assumption the general term of the series $1+8+27+64+$ etc., would be $(x+1)^3$ and not x^3.

Although, as will be shown later, any form of general term is permissible in the theory of finite differences, it will be well for the student to adopt the form suggested above until he becomes familiar with the modifications necessary when any other form is employed.

EXERCISES

Find the general term of the following series having u_0 as the first term:

1. $1+5+9+13+17+$ etc. (Check your results.)
2. $1+4+9+16+$ etc.
3. $1+0+1+4+16+$ etc.
4. $3+3^2+3^3+3^4+$ etc.
5. $8+16+32+64+$ etc.
6. $2+7+14+29+58+$ etc. *Ans.* $2+5x+x^{(2)}+x^{(3)}$.

Expand the following functions by Newton's formula:

7. x^2+x.
8. x^2-x.
9. x^3+x^2+x+1.
10. Expand x^2+x+1 by the formula

$$u_x=u_{-k}+(x+k)\,\Delta u_{-k-1}+(x+k)^{-(-2)}\frac{\Delta^2 u_{-k-2}}{2}+(x+k)^{-(-3)}\frac{\Delta^3 u_{-k-3}}{3!}+$$

etc., for $k=1$. *Ans.* $1-2(x+1)+(x+1)(x+2)$.

15. Finite Integration: Definite Integrals.—Finite integration may be defined as the inverse of finite differencing; thus, since the finite difference of x^2 is $2x+1$, the *finite integral* of $2x+1$ is x^2. Since, however, $\Delta(x^2+C)$, where C is any constant, is also $2x+1$, the finite integral of $2x+1$, written $\Sigma(2x+1)$ should be written

$$\Sigma(2x+1)=x^2+C.$$

The constant C is called the *constant of integration.* Since, in general, the value of the constant of integration is unknown, such integrals are called *indefinite integrals.*

If each of two constants is substituted for the variable in an indefinite integral and the difference between the two results taken, the constant of integration is eliminated. Such a difference is called a *definite integral* and the constants which are substituted are called *limits.* Thus, the definite integral of $2x+1$ for the limits 1 and 4 is

$$\Sigma_1^4(2x+1)=(x^2+C)_1^4=16-1=15.$$

In this case the " 4 " is called the upper limit and the " 1 " the lower limit. What would be the effect of interchanging the limits?

It is easy to verify the following fundamental formulas by differencing the expressions on the right:

$$\Sigma x^{(n)}=\frac{x^{(n+1)}}{n+1}+C. \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

$$\Sigma x^{(-n)}=\frac{x^{(-n+1)}}{-n+1}+C. \quad . \quad . \quad . \quad . \quad . \quad (6)$$

$$\Sigma ka^{mx+b}=k\frac{a^{mx+b}}{a^m-1}+C. \quad . \quad . \quad . \quad . \quad . \quad (7)$$

16. Summation of Series.—From the following scheme

$$\begin{array}{ll} u_0 & \\ & \Delta u_0 \\ u_1 & \\ & \Delta u_1 \\ u_2 & \\ & \Delta u_2 \\ \cdots & \\ & \cdots \\ u_{n-1} & \\ & \Delta u_{n-1} \\ u_n & \end{array}$$

it is evident that

$$\Delta u_0+\Delta u_1+\Delta u_2+ \ldots +\Delta u_{n-1}=\sum_{x=0}^{n-1}\Delta u_x=\Sigma_0^n\Delta u_x$$
$$=(u_x+C)_0^n=u_n-u_0. \quad . \quad (8)$$

It is very important that the distinction between the two symbols

$$\sum_{x=0}^{n-1}\Delta u_x \quad \text{and} \quad \Sigma_0^n\Delta u_x,$$

as used here, be well understood. The former means simply the sum of all terms of the form Δu_x from $x=0$ to $x=n-1$ inclusive, while the latter means the definite integral of Δu_x between the limits 0 and n.

Since u_x+C is the integral of Δu_x, relations (8) show that *the sum of a finite number of terms of a given series is given by the definite integral of the general term.* The lower limit is the same as the value of x corresponding to the first term, while the upper limit is one unit greater than the value of x corresponding to the last term. Thus, the sum of the first n terms of the series $1+3+5+7+$ etc., is

$$\sum_{x=0}^{n-1}(2x+1)=\Sigma_0^n(2x+1)=(x^2+C)_0^n=n^2.$$

It is also easily shown that

$$\sum_{x=k}^{n-1}\Delta u_x=\Sigma_k^n\Delta u_x=u_n-u_k. \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

Relation (9) is to be used when the form of the general term has not been selected in accordance with the suggestion made previously—that u_0 be the first term. Thus, the sum of the first n terms of the series considered above is given also by

$$\sum_{x=1}^{n}(2x-1)=\Sigma_1^{n+1}(2x-1),$$

and also by

$$\sum_{x=2}^{n+1}(2x-3)=\Sigma_2^{n+2}(2x-3).$$

It is suggested that, even though the sum of a specific number of terms of a series is to be determined, it is well to determine first the sum of a general number, say n, of the terms, in order that a check upon the work may be obtained; the specific number can then be substituted for n. The check is made by substituting 1, 2, etc., in turn for n in the general result found and comparing the successive results with the actual sum of one, two, etc., terms.

17. Series Whose General Terms are Rational Integral Functions.—It has already been shown that if the general term of a series is rational integral it may be easily determined by Newton's formula. Moreover, the use of Newton's formula insures that the general term so found shall be expressed in a form which can be easily integrated. Thus, for example, even though the general term of a series is known to be x^2+x+1, it would be highly inconvenient or difficult to determine the finite integral directly, but if the first few terms of the series $1+3+7+13+$ etc., were differenced and the leading term and differences were substituted in Newton's formula we obtain the general term in the form $1+2x+x^{(2)}$ which can be integrated term by term. It is easily verified that the sum of the first n terms of the series is $\frac{1}{3}(n^3+2n)$.

If only the series is given, the general term may be found, of course, in exactly the same way, the general term being assumed to be rational integral if higher differences are found to vanish.

EXERCISES

Find the sum of the first n terms of the following series:

1. $1+3+9+19+51+$ etc.

2. $-1+0+7+26+63+124+$ etc.

3. $1+3+7+13+21+31+$ etc.

4. $1+2+3+7+13+21+$ etc.

5. $1+3+6+$ etc. On what assumption can the sum obtained be regarded as correct? In what way does this example differ from the preceding examples?

6. (*a*) $1+3+5+7+$ etc.

(*b*) $1+3+5+7+10+16+28+$ etc.

Explain the difference between these two examples.

7. $2+6+18+54+162+$ etc., using $2\cdot 3^{x-1}$ as the general term.

8. $2+4+8+16+32+$ etc., using u_2 as the first term.

9. $3+6+12+24+48+$ etc.

10. $3^4+3^5+3^6+3^7+$ etc., using 3^x as the general term.

11. $1+2+5+10+17+$ etc., with u_1 as the first term.

12. $2\cdot 4+4\cdot 6+6\cdot 8+8\cdot 10+$ etc., in two ways.

(Suggestion: the general term may be written $(2x+4)^{(2)}$.)

13. $1\cdot 2+2\cdot 3+3\cdot 4+4\cdot 5+$ etc., in two ways. $(u_x=(x+2)^{(2)}.)$

14. $1\cdot 2\cdot 3+2\cdot 3\cdot 4+3\cdot 4\cdot 5+$ etc., in two ways.

15. $1\cdot 2+3\cdot 4+4\cdot 5+5\cdot 6+$ etc.

16. Show by actual summation that both $\sum_0^{n-1}(2x+1)$ and $\sum_1^{n}(2x-1)$ will give the sum of n terms of the series $1+3+5+7+$ etc.

17. Show by actual summation that both $\sum_0^{n-1}(x+1)^3$ and $\sum_1^{n}x^3$ will give the sum of n terms of the series $1+8+27+64+125+$ etc.

18. Series Whose General Terms Are Not Rational Integral.—Series whose general terms are rational integral occur quite frequently in practice but constitute only one of the large number of types which should be considered in a complete discussion of the subject of summation of series. As a complete treatment of the subject is clearly beyond the scope of this book, we shall restrict further discussion to a consideration of a few of the most important types. The general terms of these types of series will, in general, be determined by mere inspection.

Geometric series or series whose general terms are of the form ka^{mx} are readily summed by formula (7). Thus, the sum of the first n terms of the series

$$2+2\cdot 3+2\cdot 3^2+2\cdot 3^3+\text{ etc.},$$

is

$$\sum_{x=0}^{n-1} 2\cdot 3^x=(3^x)_0^n=3^n-1.$$

Many series can be summed by employing the formula known as the formula for *integration by parts*. Thus, since

$$\Delta u_x v_x = v_{x+1}\Delta u_x + u_x \Delta v_x,$$

$$\Sigma u_x \Delta v_x = u_x v_x - \Sigma v_{x+1}\Delta u_x + C. \quad . \quad . \quad . \quad . \quad (10)$$

Formula (10) is especially useful in summing various types of series whose general terms consist only in part of rational integral functions. In applying the formula it is customary to choose for u_x that part of the general term which is rational integral, so that it will appear either as a constant or with lower degree in the integral on the right. In the latter case the application of the formula must be repeated until the rational integral portion does appear as a constant on the right. The part chosen for Δv_x must, of course, be integrable. As an illustration, let us determine the sum of the first n terms of the series

$$1\cdot 1+3\cdot 3+5\cdot 3^2+7\cdot 3^3+ \text{ etc.},$$

whose general term is evidently $(2x+1)3^x$.

Letting $u_x = 2x+1$ and $\Delta v_x = 3^x$, and applying formula (10),

$$\Sigma(2x+1)3^x = (2x+1)\tfrac{3}{2}^x - \Sigma 3^{x+1} + C$$

$$= (2x+1)\tfrac{3}{2}^x - \tfrac{3}{2}^{x+1} + C.$$

Substituting the limits 0 and n and subtracting, the sum of the first n terms reduces to $3^n(n-1)+1$.

It is well to emphasize the fact that the subscript of v_x changes from x to $x+1$ each time formula (10) is applied.

Occasionally a general term is met which can be integrated either by formula (6) or by one very similar to it, but it is very important to note that in such a case the degree of the denominator must always exceed the degree of the numerator by at least 2. The difficulty otherwise encountered probably constitutes the greatest obstacle met in summing series by finite

integration. As an example illustrating the possible difficulty, let us attempt to determine the sum of the first n terms of the series

$$1+\tfrac{1}{2}+\tfrac{1}{3}+\tfrac{1}{4}+ \text{ etc.}$$

The general term is $1/x+1$ or $(x+1)^{(-1)}$. According to Exercise 7, p. 18.

$$\Sigma(x+1)^{(-1)}=\frac{(x+1)^0}{0}+C,$$

which is meaningless. Interesting attempts have been made to overcome such difficulties, and interesting results have been attained, but as the results are of little practical value for present purposes further consideration of the problem must be omitted here.

One type of a rational function (defined as the quotient of two rational integral functions) has a general term of the form $f(x)(x+k)^{(-n)}$ or more rarely $f(x)(ax+b)^{(-n)}$ where $f(x)$ is rational integral. In either case it is only necessary to expand the numerator into a series with terms of the same form as those appearing in the denominator, and then to break up the general term so expressed into a sum of several expressions. The general expression for the numerator can be determined by inspection or by Newton's formula, and its expansion can usually be made by inspection. As an example, the general term of the series

$$\frac{1}{1\cdot2\cdot3}+\frac{3}{2\cdot3\cdot4}+\frac{5}{3\cdot4\cdot5}+\frac{7}{4\cdot5\cdot6}+ \text{ etc.},$$

is evidently

$$(2x+1)(x+1)^{(-3)}=\frac{2x+1}{(x+1)(x+2)(x+3)},$$

and the numerator $2x+1$ can be written by inspection $-1+2(x+1)$. (See Exercise 10, p. 21.) Hence, the general term can be broken up into two parts $-(x+1)^{(-3)}+2(x+2)^{(-2)}$.

It is easily verified that the sum of n terms of the series is,

$$\frac{1}{2(n+1)(n+2)}-\frac{2}{n+2}+\frac{3}{4}.$$

(See Exercise 7, p. 18.)

EXERCISES

Find the sum of n terms of the series:

1. $-1\cdot1+2\cdot2+5\cdot2^2+8\cdot2^3+\ldots+(3x-1)2^x+$ etc.
2. $1\cdot1+2\cdot3+5\cdot3^2+10\cdot3^3+\ldots+(x^2+1)3^x$, etc.
3. Whose general term is $(2x+3)(x+2)^{(-3)}$.
4. Whose general term is $(3x+4)(x-1)^{(-3)}$.
5. Whose general term is $\dfrac{2x+1}{(x+1)(x+3)(x+4)}$. Multiply numerator and denominator by $x+2$.
6. Whose general term is $\dfrac{2x+2}{(3x+1)(3x+4)(3x+7)}$. Use the formula

$$\Sigma(ax+b)^{(n)}=\frac{(ax+b)^{(n+1)}}{a(n+1)} \quad \text{with } n \text{ negative.}$$

MISCELLANEOUS EXERCISES

Determine the sum of the first n terms of the following series:

1. $1+4+9+16+25+\ldots.$ *Ans.* $\frac{1}{6}(2n^3+3n^2+n)$.
2. $1+5+11+19+29+\ldots.$
3. $1+8+27+64+125+\ldots.$
4. $1+5+5^2+5^3+5^4+\ldots.$
5. $6^3+6^4+6^5+6^6+\ldots.$
6. $3+8+15+24+35+\ldots.$
7. $2+12+36+80+150+\ldots.$
8. $4+18+48+100+180+\ldots.$
9. $1\cdot1+2^2\cdot2+3^2\cdot2^3+4^2\cdot2^3+\ldots.$
10. $1\cdot2+2^2\cdot2^2+3^2\cdot2^3+4^2\cdot2^4+\ldots.$
11. $1\cdot1+3\cdot5+5\cdot5^2+7\cdot5^3+9\cdot5^4+\ldots.$
12. $1\cdot3+4\cdot3^2+9\cdot3^3+16\cdot3^4+25\cdot3^5+\ldots.$

13. $\frac{1}{1\cdot 2}+\frac{1}{2\cdot 3}+\frac{1}{3\cdot 4}+\frac{1}{4\cdot 5}+\ldots.$

14. $\frac{-1}{3\cdot 4\cdot 5}+\frac{1}{4\cdot 5\cdot 6}+\frac{3}{5\cdot 6\cdot 7}+\frac{5}{6\cdot 7\cdot 8}+\ldots.$

15. $\frac{1}{1\cdot 3}+\frac{1}{2\cdot 4}+\frac{1}{3\cdot 5}+\ldots.$

16.[1] $\frac{1}{1\cdot 3\cdot 5}+\frac{4}{3\cdot 5\cdot 7}+\frac{7}{5\cdot 7\cdot 9}+\frac{10}{7\cdot 9\cdot 11}+\ldots.$

17. (*a*) $1+3+6+10+15+\ldots.$

(*b*) $1+3+6+10+15+22+33+51+\ldots.$

Compare the results obtained in (*a*) and (*b*). Explain the nature of the assumptions underlying each problem.

18. $0+0+0+6+24+60+120+\ldots.$

(*a*) using the first three terms.

(*b*) omitting the first three terms. Compare the results and explain the difference.

19. Algebraic Treatment of Symbols.—Let us now refer back to some of the operations of the finite calculus and introduce the derivative to arrive at an interesting formula.

If we define the *operator* E by the relation

$$Eu_x = u_{x+1},$$

so that

$$E^n u_x = u_{x+n},$$

then, since

$\Delta^n u_x = u_{x+n} - nu_{x+n-1} + \frac{n(n-1)}{2} u_{x+n-2} -$ etc. (see Ex. 7b, p. 16),

we may write

$$\Delta^n u_x = \left(E^n - nE^{n-1} + \frac{n(n-1)}{2} E^{n-2} - \text{etc.}\right) u_x,$$

where all the terms within the parenthesis are to be considered

[1] The general expression for the numerator can be obtained by Newton's formula.

as *operating* upon u_x; or, expressing the relation more compactly,

$$\Delta^n u_x = (E-1)^n u_x,$$

or

$$\Delta^n = (E-1)^n,$$

and

$$\Delta = E-1. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

It is left for the student to show that Newton's formula can be expressed symbolically,

$$E^n u_x = (1+\Delta)^n u_x \text{ (see Ex. 7a, p. 16),}$$

or

$$E = 1+\Delta,$$

which agrees with (11). The preceding is a good illustration of the possibilities of treating various operations as algebraic operations. Although such a procedure calls for special interpretation and a certain amount of check in some cases, as we shall find, it frequently enables one to arrive at results with much less labor than the ordinary procedure would require.

Another illustration of the algebraic treatment of symbols, which is useful, is as follows:

Taylor's expansion of the ordinary calculus may be written

$$f(x+h) = f(x) + hD_x f(x) + \frac{h^2}{2} D_x{}^2 f(x) + \frac{h^3}{3} D_x{}^3 f(x) + \text{ etc.},$$

where D_x is employed as the symbol for the derivative.

If we let $h=1$ and use the notation peculiar to finite differences

$$u_{x+1} = Eu_x = u_x + D_x u_x + \frac{D_x{}^2 u_x}{2} + \frac{D_x{}^3 u_x}{3} + \text{ etc.}$$

$$= (1 + D_x + \frac{D_x{}^2}{2} +, \text{ etc.}) u_x$$

$$= e^{D_x} u_x \text{ (see the expansion of } e^x \text{ given below)}$$

or

$$E = e^{D_x}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

Thus, E, Δ, and e^{Dx} are seen to be connected symbolically by the relations

$$E=1+\Delta=e^{Dx}. \quad . \quad . \quad . \quad . \quad . \quad . \quad (13)$$

20. Bernoulli's Numbers.—The coefficients of the terms of the expansion

$$\frac{x}{e^x-1}=B_0-B_1x+B_2\frac{x^2}{2}-B_4\frac{x^4}{4!}+B_6\frac{x^6}{6!}-\text{ etc.}, \quad . \quad (14)$$

or B_0, B_1, B_2, etc., are known as Bernoulli's Numbers. The expansion is easily obtained by ordinary division, where e^x is replaced by the series

$$1+x+\frac{x^2}{2}+\frac{x^3}{3!}+\frac{x^4}{4!}+\text{ etc.}$$

The values of a few of the numbers are as follows:

$$B_0=1 \qquad B_4=\tfrac{1}{30}$$
$$B_1=\tfrac{1}{2} \qquad B_6=\tfrac{1}{42}$$
$$B_2=\tfrac{1}{6} \qquad \text{etc.}$$

All the numbers corresponding to odd subscripts are zero, except B_1; this fact can be verified by inspection of the following identity,

$$\frac{x}{e^x-1}+\frac{x}{2}=\frac{x}{2}\cdot\frac{e^x+1}{e^x-1},$$

for the left side is the expansion (14) minus the second term, and if the sign of x is changed on the right side the expression remains unaffected, showing that it is an *even* function or that (14) contains no odd powers after the second term.

21. Summation of Series.—The expansion (14) given in the preceding section proves to be well adapted, with certain interpretations, to summation of series. It has been employed to sum series of certain troublesome forms to give approximate results, but we shall have to omit the consideration of these special forms here.

Since $\Delta = e^{D_x} - 1$,

$$\Sigma u_x = \Delta^{-1} u_x = (e^{D_x} - 1)^{-1} u_x = \frac{u_x}{e^{D_x} + 1}$$

$$= B_0 D_x^{-1} u_x - B_1 u_x + B_2 \frac{D_x u_x}{2} - B_4 \frac{D_x^3 u_x}{4!} + \ldots$$

where, it will be noted, the various powers of D_x are interpreted as orders of differentiation. If, in addition, we interpret $D_x^{-1} u_x$ as the inverse of differentiation, or as integration, we obtain finally

$$\Sigma u_x = C + \int u_x dx - \frac{u_x}{2} + \frac{D_x u_x}{12} - \frac{D_x^3 u_x}{720} + \frac{D_x^5 u_x}{30240} - \text{ etc.} \quad . \quad (15)$$

As an application of formula (15) let us find the sum of the first n terms of the series $0^4 + 1^4 + 2^4 + 3^4 +$ etc.

It will be helpful to verify the following results:

$$\sum_0^{n-1} x^4 = \left(C + \int x^4 dx - \frac{x^4}{2} + \frac{D_x x^4}{12} - \frac{D_x^3 x^4}{720} \right)_0^n$$

$$= \frac{n^5}{5} - \frac{n^4}{2} + \frac{n^3}{3} - \frac{n}{30}.$$

Verify the result for the sum of 1, 2 and 3 terms respectively.

EXERCISES

Find the sum of the first n terms of the series whose general terms are:

1. $x + 3$. (Check your results.)
2. $x^2 + 2$.
3. $x^3 + 1$.

CHAPTER III

INTERPOLATION

22. Interpolation.—Even though the functional relation between two or more variables is given, it may prove very troublesome to evaluate the function for given values of the independent variables. Take, for example, the equation $y = \log_{10} x$. Although it is not particularly difficult to compute the logarithm of a given number, say by the use of an infinite series, such a method requires too much labor and time to be employed in an ordinary application of logarithms. The method ordinarily employed is to have a table of the most frequently desired values of such a function and then to obtain other values by proportion. The observing student will note that such an assumption amounts to an assumption that the graph of the function is a straight line. Although such an assumption is clearly unjustified with respect to the graph of any function taken as a whole, it proves quite reasonable in many cases when applied to the values of the function within a relatively small interval, and values so obtained often prove sufficiently accurate to meet the needs at the time; the process then avoids much labor in computation. The scheme just considered is familiar to the student and is already known to him as interpolation. We shall use the term *interpolation* in the same connection; but we shall extend the application of the scheme to cases where the values of a function within a given interval are assumed to satisfy a more general function, namely, a rational integral function or polynomial of the form $y = a + bx + cx^2 +$ etc., sufficiently accurately to meet the immediate needs. As we have already found—by differencing—there

are many functions which will not justify such an assumption at all; on the other hand, there is a surprisingly large number of functions which lend themselves much more satisfactorily to the plan we propose to consider than the nature of the function itself would probably indicate. The final test as to whether the plan is feasible in any particular case will lie in the relative sizes of the successive differences of the given values. This fact extends the usefulness of the scheme, because in a great many cases only the numerical values will be given and nothing of the nature of the fundamental function will be known.

We shall restrict our attention to the interpolation of values of functions of a single variable.

23. Newton's Formula.—Newton's formula proves admirably suited for interpolation from the point of view of simplicity and ease in computation. Primarily, Newton's formula represents a curve which passes through the points $(0, u_0)$, $(1, u_1)$, etc., whose equation is rational integral and in general of the $(n-1)$th degree if there are n points. As a concrete illustration, the output of steel in the United States in hundreds of thousands of tons for the four specified years was:

1890	$u_0 = 43$			
		$\Delta u_0 = 18$		
1895	$u_1 = 61$		$\Delta^2 u_0 = 23$	
		41		$\Delta^3 u_0 = 34$
1900	$u_2 = 102$		57	
		98		
1905	$u_3 = 200$			

Hence, the estimated output for, say, year 1897 would be $u_{7\frac{1}{5}}$ obtained from Newton's formula in the form

$$u_x = 43 + 18x + 23\frac{x(x-1)}{2} + 34\frac{x(x-1)(x-2)}{6},$$

where the interpolation is restricted to third differences, or

$$u_{7\frac{1}{5}} = 43 + 18 \cdot \tfrac{7}{5} + 23 \cdot \tfrac{7}{2\,5} - 34 \cdot \tfrac{7}{1\,2\,5}$$

$$= 72.736 \text{ or } 73.$$

What would have been the result of restricting the work to second differences? To first differences? Restricting the use of Newton's formula to first, second, third, etc., differences amounts obviously to passing curves of the first (straight line), second (parabola), third, etc., degree respectively through two, three, four, etc., points respectively.

It should be emphasized that u_0 can be assigned at will to whatever given value we please, and u_1, u_2, etc., in accordance with this assignment; once, however, the assignment is made, the abscissa of the ordinate or value to be interpolated is uniquely but easily determined by inspection; moreover, the use of Newton's formula in the form given above (that is, in terms of u_0 and its differences) requires that the first value given shall be u_0 or the value corresponding to $x=0$.

The most prominent statisticians have come to agree that third or fourth differences are usually sufficient when dealing with ordinary statistical data; this fact is kept in mind in the treatment and the illustrations given in the following pages.

Attention is called to the remarkable fact that, although the logarithmic function is irrational by nature and therefore has no order of differences that vanishes *absolutely*, differences of numerical values of the function taken from a reasonably small interval converge so rapidly in practice that it usually requires few differences to interpolate with a remarkable degree of accuracy. As an example, suppose that we have given the following values:

$$\log 2.7182 = 0.4342814081$$

$$\log 2.7183 = 0.4342973851$$

$$\log 2.7184 = 0.4343133615$$

$$\log 2.7185 = 0.4343293373$$

$$\log 2.7186 = 0.4343453126$$

If we difference these values and employ Newton's formula, we obtain

$$\log e = 0.4342814081 + 0.8182846(159770 \cdot 10^{-10})$$

$$+ \frac{0.82(0.18)}{2}(6 \cdot 10^{-10})$$

$$= 0.4342944819 \text{ (correct to the last place)}$$

where $e = 2.71828182846$. The values of x and $1-x$ are reduced to two decimals in multiplying the second difference because of the relative smallness of the latter.

24. Lagrange's Formula: Central Difference Formulas.— If the series of given values are not " equidistant " a convenient formula to be used is that known as Lagrange's formula (although the formula is applicable also when the values are equidistant). This formula, however, is not based upon finite differences and proves rather cumbersome in application; its selection, therefore, is usually a matter of necessity rather than of preference. If the n values $f(a), f(b), f(c), \ldots f(k)$ are given, we assume a rational integral function of the $(n-1)$th degree of the following form:

$$f(x) = A(x-b)(x-c)\ldots(x-k) + B(x-a)(x-c)\ldots(x-k) + \text{etc.},$$

where a different one of the factors $(x-a)$, $(x-b)$, etc., is missing in each product.

Substituting $x = a$ we obtain

$$A = \frac{f(a)}{(a-b)(a-c)\ldots(a-k)},$$

likewise, for $x = b$

$$B = \frac{f(b)}{(b-a)(b-c)\ldots(b-k)},$$

and so on for other coefficients. Substituting these expressions in the equation assumed originally, we obtain

$$f(x)=f(a)\frac{(x-b)(x-c)\ \ldots\ (x-k)}{(a-b)(a-c)\ \ldots\ (a-k)}$$

$$+f(b)\frac{(x-a)(x-c)\ \ldots\ (x-k)}{(b-a)(b-c)\ \ldots\ (b-k)}+\text{ etc.} \quad . \quad (16)$$

which is known as Lagrange's formula.

As an illustration, suppose that we select the values of the output of steel for the years 1890, 1900 and 1905 given in the preceding section; then we may let $a=0$, $b=2$, and $c=3$, and $f(0)=43$, $f(2)=102$ and $f(3)=200$; formula (16) then becomes

$$f(x)=43\frac{(x-2)(x-3)}{(-2)(-3)}+102\frac{x(x-3)}{(2)(2-3)}+200\frac{x(x-2)}{3(3-2)}.$$

It is easily verified that if we substitute $7/5$ for x in this equation we obtain 65 as the estimated output for 1897. It is left for the student to show that if formula (16) were applied to all the data of the original problem the result would be the same as that obtained previously. Why must the result be the same?

Sometimes the data are known to be very accurate and accuracy in the final results of interpolation is very essential. Under such circumstances it is well to consider a certain fault in Newton's formula and employ a suitable modification of it. Newton's formula is expressed, of course, in terms of u_0 and its differences; since u_0 is an "end" value it receives more emphasis than it deserves. Slightly better results would be obtained if the interpolations were made in a central interval in terms of values located on both sides of the interpolated values, without giving too much emphasis to the values on any one side. Formulas have been derived with this idea in mind and are called *central difference* formulas; all can be derived by simple modifications of Newton's formula, but as they would be valuable only in connection with data which would be more accurate than ordinary statistical data we shall omit further consideration of them here. The student who is contemplat-

ing work of a more refined character should, however, keep them in mind.

EXERCISES

1. Find the logarithms of the following numbers (using the tables given in the back of this book):

(*a*) 93.4632
(*b*) 0.632481
(*c*) 0.00128734

(The final results are likely to prove a unit or so in error in the last place. At what part of the table are the results more likely to be in error?)

2. Find the antilogarithms of the following logarithms (using the tables given in this book):

(*a*) 2.834672.
(*b*) 4.164872.
(*c*) 8.426838−10.

3. Inverse interpolation consists of solving backward for x. Use Newton's formula and find the antilogarithm of the following numbers, *using the values given in the table of logarithms*:

(*a*) 2.368427, using first differences only;
(*b*) 2.368427, using second differences and solving a quadratic equation.

4. Find the logarithm of 132 from the logarithms of 131, 133 and 135 (using the tables given in this book):

(*a*) by Lagrange's formula;
(*b*) by Newton's formula.

5. Test the following sets of values as to whether satisfactory interpolation by Newton's formula could be expected:

(*a*) The amounts to which \$1 would accumulate at compound interest at 2 per cent in 16, 17, etc., years:

Years	Amount
16	1.37278571
17	1.40024142
18	1.42824625
19	1.45681117
20	1.48594740

(*b*) The reciprocals of the numbers:

$7651 = .0001307019\ldots$

$7652 = .0001306849\ldots$

$7653 = .0001306677\ldots$

$7654 = .0001306506\ldots$

$7655 = .0001306336\ldots$

6. The pressure of wind in pounds per square foot corresponding to the velocity in miles per hour has been determined by experiment to be approximately as follows:

Velocity	Pressure
15	1.1
20	2.0
30	4.4
40	7.9

Estimate the pressure for a velocity of 25 miles per hour.

7. Given $\log 71 = 1.8512583$
$\log 72 = 1.8573325$
$\log 73 = 1.8633229$
$\log 74 = 1.8692317$

find $\log 71.54$. Find also $\log 0.07154$.

8. Given $\log \sin 12\ 39' = 9.34043382 - 10$
$\log \sin 12\ 40' = 9.34099630 - 10$
$\log \sin 12\ 41' = 9.34155802 - 10$
$\log \sin 12\ 42' = 9.34211897 - 10$
$\log \sin 12\ 43' = 9.34267917 - 10$

find log sin 12 40′.4134. (The computation may be simplified by rounding the values of x, $x-1$, etc., consistent with the relative values of the differences.) *Ans.* 9.34022108−10.

9. Given log cos 27 36′=9.94753370−10
log cos 27 37′=9.94746743−10
log cos 27 38′=9.94740132−10
log cos 27 39′=9.94733516−10
log cos 27 40′=9.94726895−10

find log cos 27 38′.3. *Ans.* 9.94738147−10.

10. Given log 472=2.6739419986
log 473=2.6748611407
log 474=2.6757783417
log 475=2.6766936096
log 476=2.6776069527,

find log 472832.

11. Given the death rates per 100,000 population in the registration area of the United States by years for the following diseases:

	Typhoid	Tuberculosis	Cancer
1906	31.3	157.1	69.1
1909	21.1	139.3	73.8
1912	16.5	129.8	77.0
1915	12.4	127.7	81.1

estimate the death rates for the year 1910 by Newton's formula.

25. Leading-difference Formulas.—If any formula based upon finite differences is to be employed to interpolate several values in the same interval, the work can be simplified and systematized somewhat further by formulas for the leading differences of these interpolated values. As an illustration we shall show how Newton's formula can be so applied to interpolate $t-1$ values in a central interval. If we difference the

values of $u_{\underset{t}{t}}$, $u_{\underset{t}{t+1}}$, etc., obtained from Newton's formula (to third differences) three times, the leading differences for interpolating $t-1$ values by third differences between u_1 and u_2 become

$$(1) = \frac{\Delta u_0}{t} + \frac{t+1}{2}\frac{\Delta^2 u_0}{t^2} - \frac{t^2-1}{6}\frac{\Delta^3 u_0}{t^3}$$

$$(2) = \quad 1\frac{\Delta^2 u_0}{t^2} + \quad 1\frac{\Delta^3 u_0}{t^3}. \quad . \quad . \quad (17)$$

$$(3) = \quad 1\frac{\Delta^3 u_0}{t^3}$$

$$(4) = (5) = \text{etc.}, = 0$$

For $t=5$ these leading differences become

$$(1) = .2\Delta u_0 + .12\Delta^2 u_0 - .032\Delta^3 u_0$$

$$(2) = \quad .04\Delta^2 u_0 + .008\Delta^3 u_0. \quad . \quad . \quad (17')$$

$$(3) = \quad .008\Delta^3 u_0$$

As an example, suppose we wish to interpolate four equidistant values between u_1 and u_2 of the following hypothetical set of values:

$u_0 = 416$			
	138		
$u_1 = 554$		-34	
	104		8
$u_2 = 658$		-26	
	78		
$u_3 = 736$			

Then by (17')

$$(1) = .2(138) - 3(.04)34 - 4(.008)8 = 23.264$$

$$(2) = -1.296$$

$$(3) = 0.064$$

The differences just found are then added accumulatively to $u_1 = 554$, as follows, to give the desired interpolations:

$u_1 = 554.000$			
	$(1) = 23.264$		
577.264		$(2) = -1.296$	
	21.968		$(3) = .064$
599.232		-1.232	
	20.736		.064
619.968		-1.168	
	19.568		.064
639.536		-1.104	
	18.464		
$u_2 = 658.000$			

Probably the main advantage to be gained by the use of such formulas is the check upon the work; for not only are the $t-1$ (4 in the example) values interpolated, say between u_1 and u_2, but the "end" value u_2 is also reproduced—if no error is made in the computation. The work was carried to the last decimal place in the above example to show this check; it would be unnecessary to retain all the decimal places in practice.

26. Tangential Interpolation.—If the method of interpolating several values in an interval, explained in the preceding section, is applied to several succeeding intervals, the interpolation curve passing through the values interpolated between, say, y_1 and y_2 will not in general be continuous with the curve passing through the values interpolated between y_2 and y_3, etc. Hence, in the final series of interpolated values there will in general be discontinuities at y_2, y_3, y_4, etc. It is possible to adjust whatever interpolation formula is employed so that any two interpolation curves will have the same slope at the point of their intersection, that is, at the point which constitutes the "end" point of one interval and the "beginning" point of the next interval. Interpolation based upon such a scheme is called *tangential*[1] interpolation.

[1] If two interpolated curves are required to have not only the same slopes but also the same curvatures at their points of intersection, the

The scheme will now be considered in connection with Newton's formula.

We shall assume the tangential formula to be of the form

$$y_x = a + bx + cx^2 + dx^3. \quad . \quad . \quad . \quad . \quad . \quad (A)$$

and our immediate problem is to determine the coefficients a, b, c and d.

If the corresponding curve passes through y_1 and y_2 we must have

$$y_1 = a + b + c + d = y_0 + \Delta y_0. \quad . \quad . \quad . \quad . \quad . \quad . \quad (B)$$

and

$$y_2 = a + 2b + 4c + 8d = y_0 + 2\Delta y_0 + \Delta^2 y_0. \quad . \quad . \quad (C)$$

Next we require that the slope of curve (A) at y_1 shall be the same as that of the parabola through y_0, y_1, and y_2 or

$$y_x = y_0 + x\Delta y_0 + \frac{x(x-1)}{2}\Delta^2 y_0,$$

at the same point. It is easily verified that this requirement leads to the relation

$$b + 2c + 3d = \Delta y_0 + \tfrac{1}{2}\Delta^2 y_0. \quad . \quad . \quad . \quad . \quad . \quad (D)$$

Finally, we require that the slope of curve (A) at y_2 shall be the same as the slope of the parabola through y_1, y_2 and y_3 whose equation is the same as that of the parabola through y_0, y_1 and y_2, except that y_0 and its differences should be replaced by y_1 and its differences. The value of the slope at y_2 is then $\Delta y_1 + \frac{1}{2}\Delta^2 y_1$. If we express the value of this slope in terms of y_0 and its differences, this requirement leads to the relation

$$b + 4c + 12d = \Delta y_0 + \tfrac{3}{2}\Delta^2 y_0 + \tfrac{1}{2}\Delta^3 y_0. \quad . \quad . \quad . \quad (E)$$

interpolation is called *osculatory*, because two such adjacent curves are said to have a common osculating circle at such a point. Osculatory interpolation, however, will be found to require fifth differences, and as fifth differnces are ordinarily inappropriate for our purposes we shall give our sole attention to tangential interpolation, which calls for only third differences.

Solving relations (B), (C), (D) and (E) simultaneously for a, b, c and d and substituting these values in equation (A), we obtain the tangential formula

$$y_x = y_0 + x\Delta y_0 + \frac{x(x-1)}{2}\Delta^2 y_0 + \frac{(x-1)^2(x-2)}{2}\Delta^3 y_0. \quad . \quad (18)$$

If we refer to the values y_0, y_1 and y_2 for the present as " the first set " and y_1, y_2 and y_3 as " the second set " for the interval from y_1 to y_2, it should be evident that in interpolating in that interval by means of the differences of y_0, y_1, y_2 and y_3 by formula (18) the slope of the curve at y_1 is determined by the first set of values and the slope at y_2 by the second set. Likewise, in interpolating in the succeeding interval from y_2 to y_3, the slope at y_2 is determined by the first set *for that interval* which, however, is identical with the second set of the preceding interval (that is, from y_1 to y_2); hence, the interpolation curves of the two intervals must have the same slope at their point of intersection y_2; and so on for all succeeding intervals.

The leading differences of formula (18) for interpolating $t-1$ values in the interval from y_1 to y_2 are as follows:

$$\begin{aligned}
(1) &= \frac{\Delta y_0}{t} + \frac{t+1}{2}\cdot\frac{\Delta^2 y_0}{t^2} - \frac{t-1}{2}\,\frac{\Delta^3 y_0}{t^3} \\
(2) &= \quad 1 \quad \frac{\Delta^2 y_0}{t^2} - (t-3)\,\frac{\Delta^3 y_0}{t^3}. \quad . \quad . \quad . \quad (19) \\
(3) &= \quad + \quad 3\,\frac{\Delta^3 y_0}{t^3} \\
(4) &= (5) = \text{etc.}, = 0
\end{aligned}$$

For $t=5$ these leading differences become

$$\begin{aligned}
(1) &= .2\Delta y_0 + .12\Delta^2 y_0 - .016\Delta^3 y_0 \\
(2) &= \quad .04\Delta^2 y_0 - .016\Delta^3 y_0. \quad . \quad . \quad . \quad (19') \\
(3) &= \quad .024\Delta^3 y_0.
\end{aligned}$$

27. Interpolation of Ordinates Among Areas.—So far, interpolations have been concerned entirely with ordinates; that is, both the data and the values to be interpolated have been essentially ordinates. We shall now derive a useful formula to be used for interpolating ordinates where, however, the data consist essentially of areas. Suppose, for example, that it is desired to estimate the value of crude materials, for use in manufacturing silk, imported into the United States for the *single* year 1907, from the values for the following *groups* of years:

	Millions of Dollars
1900–4	1479
1905–9	2096
1910–4	2902

It is evident that no formula derived so far will enable us to determine the desired value. We shall return to this particular problem later.

Suppose that $u_{\underset{t}{x}}$ and $y_{\underset{t}{x}}$ are two functions which have the relation

$$u_{\underset{t}{x}} = \Delta y_{\underset{t}{x}} = y_{\underset{t}{x+1}} - y_{\underset{t}{x}}. \quad . \quad . \quad . \quad . \quad . \quad . \quad (A)$$

Hence

$$y_{\underset{t}{x+t}} = \sum_{x=x}^{x+t-1} u_{\underset{t}{x}} + y_{\underset{t}{x}}.$$

Also, let

$$w_{\underset{t}{x}} = u_{\underset{t}{x}} + u_{\underset{t}{x+1}} + \ldots + u_{\underset{t}{x+t-1}}.$$

Hence,

$$y_{\underset{t}{x+t}} - y_{\underset{t}{x}} = w_{\underset{t}{x}}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (B)$$

Substituting in (B)

$$\begin{array}{llll} x=0 & y_1-y_0= & \Delta y_0=w_0 & \\ & & & \Delta^2 y_0=\Delta w_0 \\ x=t & y_2-y_1= & \Delta y_1=w_1 & \Delta^3 y_0=\Delta^2 w_0 \\ & & & \Delta^2 y_1=\Delta w_1 \\ x=2t & y_3-y_2= & \Delta y_2=w_2 & \end{array}$$

etc.,

If now we expand $y_{\underline{x+1}}$ and $y_{\underline{x}}$ by Newton's formula and take the difference between the results, in accordance with relation (A), replacing Δy_0 by w_0, $\Delta^2 y_0$ by Δw_0, etc., we obtain

$$u_{\underline{x}} = \frac{w_0}{t} + (2x-t+1)\frac{\Delta w_0}{2t^2} + \{3x^2-3x(2t-1)+2t^2-3t+1\}\frac{\Delta^2 w_0}{6t^3} + . \quad (20)$$

For $t=5$ formula (20) becomes

$$u_{\underline{x}} = \frac{w_0}{5} + (x-2)\frac{\Delta w_0}{5^2} + (x^2-9x+12)\frac{\Delta^2 w_0}{2.5^3} + \ldots \quad (20')$$

Now, the values w_0, w_1, w_2, etc., are obviously groups of t values of which the values given in the problem suggested above are examples, and $u_{\underline{x}}$ represents one of the values included in such a group whose identification depends upon the value of x; in the problem given above (where $t=5$) $u_{\underline{0}}$, $u_{\underline{1}}$, etc., represent the values corresponding to the individual years 1900, 1901, etc., and $u_{\underline{0}} + u_{\underline{1}} + \ldots + u_{\underline{4}} = w_0 = 1479$, etc.

Hence,

$$\begin{array}{lll} w_0 = 1479 & & \\ & \Delta w_0 = 617 & \\ w_1 = 2096 & & \Delta^2 w_0 = 189 \\ & 806 & \\ w_2 = 2902 & & \end{array}$$

and the value corresponding to the individual year 1907 (for $x=7$) is

$$u_{\underline{7}} = .2(1479) + .2(617) - .008(189) \text{ (to second differences)}$$

$$= 417.688 \text{ or } 418.$$

The leading term (why does this case require a leading *term*?) and differences for breaking up completely the group w_1 (given w_0, w_1 and w_2) are as follows:

$$\text{Leading term} = \frac{w_0}{t} + \frac{(t+1)}{2}\frac{\Delta w_0}{t^2} + \frac{(1-t^2)}{6}\frac{\Delta^2 w_0}{t^3}$$

$$(1) = \quad 1\,\frac{\Delta w_0}{t^2} + \quad 1\,\frac{\Delta^2 w_0}{t^3}. \quad . \quad (21)$$

$$(2) = \quad 1\,\frac{\Delta^2 w_0}{t^3}$$

It is easily verified that the leading term and differences for the values given above are 363.792, 26.192 and 1.512 respectively. If these values are written and added accumulatively as follows, we obtain the values corresponding to the individual years as desired:

Years	Values (in millions of dollars)		
1905	363.792		
		26.192	
1906	389.984		1.512
		27.704	
1907	417.688		1.512
		29.216	
1908	446.904		1.512
		30.728	
1909	477.632		
	2096.000		

The work was carried to a greater number of decimal places than would be necessary in practice, in order to show a perfect check; that is, to show that the sum of the values for the individual years equals $w_1 = 2096$.

EXERCISES

1. Given the consumption of coffee in the United States for the quinquennial periods:

1895– 9	50.03 pounds per capita
1900– 4	55.90
1905– 9	54.20
1910–14	46.75,

estimate the consumption per capita for the year 1902.

2. From the data of problem (1), estimate the consumption per capita for the individual years 1900, 1901, . . . 1904.

3. From the number of patents (in thousands) issued in the United States for the years:

1895– 9	117
1900– 4	144
1905– 9	170
1910–14	185,

estimate the number of patents for the years 1900, 1901, . . . 1904.

4. The enrollment (men and women) in the colleges, universities, and schools of technology of the United States was as follows:

1895– 9	437 thousands
1900– 4	530
1905– 9	691
1910–14	963

Estimate the enrollment for the individual years 1900, 1901, . . . 1904.

5. The number of American and Filipino teachers in the Philippine Islands were as follows:

Years Beginning	American	Filipino
1904– 6	2432	14,896
1907– 9	2357	23,028
1910–12	2005	23,112
1913–15	1638	28,371

Estimate the number of (*a*) American and of (*b*) Filipino teachers for the individual years 1907, 1908 and 1909. Estimate the number of (*c*) American and of (*d*) Filipino teachers for the year 1910.

28. Areas from Representative Ordinates.—Sometimes the problem is reversed in the sense that every t-th individual value is known and it is desired to determine the values corresponding to each group. As in the previous case, second differences will probably prove sufficient for ordinary purposes.

If we substitute $\frac{t}{t}, \frac{t+1}{t}, \ldots \frac{2t}{t}$ in Newton's formula we obtain respectively

$$u_{\frac{t}{t}} = u_0 + t\frac{\Delta u_0}{t}.$$

$$u_{\frac{t+1}{t}} = u_0 + (t+1)\frac{\Delta u_0}{t} + (1+t)\frac{\Delta^2 u_0}{2t^2}.$$

$$u_{\frac{t+2}{t}} = u_0 + (t+2)\frac{\Delta u_0}{t} + 2(2+t)\frac{\Delta^2 u_0}{2t^2}.$$

$$u_{\frac{t+3}{3}} = u_0 + (t+3)\frac{\Delta u_0}{t} + 3(3+t)\frac{\Delta^2 u_0}{2t^2}.$$

$$u_{\frac{t+t-1}{t}} = u_0 + (t+t-1)\frac{\Delta u_0}{t} + (t-1)(t-1+t)\frac{\Delta^2 u_0}{2t^2}.$$

$$u_{\frac{2t}{t}} = u_0 + 2t\frac{\Delta u_0}{t} + t(t+t)\frac{\Delta^2 u_0}{2t^2}$$

The sum of the first t lines gives $w_{\frac{t}{t}}$ and the sum of the last t lines gives $w_{\frac{t+1}{t}}$ or

$$w_{\frac{t}{t}} = tu_0 + \frac{3t-1}{2}\Delta u_0 + \tfrac{1}{6}(5t^2-6t+1)\frac{\Delta^2 u_0}{2t} + \text{etc.} \quad . \quad (22)$$

$$w_{\frac{t+1}{t}} = tu_0 + \frac{3t+1}{2}\Delta u_0 + \tfrac{1}{6}(5t^2+6t+1)\frac{\Delta^2 u_0}{2t} + \text{etc.} \quad . \quad (23)$$

Formula (22) is called an *initial* form and formula (23) a *terminal* form, for obvious reasons; their use is made clear by the following illustrations: The values of the products of silk manufactures (in thousands of dollars) of the United States are given by a certain authority only for individual years; particular values are as follows:

Year	Values (in thousands of dollars)
1899	107,256
1904	133,288
1909	196,912
1914	254,011

Suppose that it is desired to estimate the total products for the quinquennial period 1905–9. It is evident that the terminal form of the formulas derived above is needed. Differencing the given values and substituting in formula (23), we obtain

$$w_{\frac{6}{5}} = 5(107{,}256) + 8(26{,}032) + 2.6(37{,}592)$$
$$= 842{,}275.$$

If the values were given for the years 1900, 1905, etc., the initial form would obviously be required.

In order to interpolate the values for "end" groups (for example, for 1910–14 or 1900–4) it is necessary to use one of the formulas

$$w_{\frac{0}{t}} = tu_0 + \frac{t-1}{2}\Delta u_0 + \frac{1-t^2}{12t}\Delta^2 u_0. \quad . \quad . \quad . \quad (24)$$

$$w_{\frac{1}{t}} = tu_0 + \frac{t+1}{2}\Delta u_0 + \frac{1-t^2}{12t}\Delta^2 u_0, \quad . \quad . \quad . \quad (25)$$

where (24) is obtained by adding the first t values and (25) by adding the last t values obtained by substituting $0/t$, $1/t$, . . . t/t in Newton's formula. It should be noticed, however, that for the data of this problem the total value of the products for the period 1910–14 would be obtained by reversing the order of the given values (before differencing) and applying the initial form (24).

EXERCISES

1. The value of silk products (in thousands of dollars) in the United States, previous to 1904, were known only for the years:

1869	12,211
1879	41,033
1889	87,298
1899	107,256]

Estimate the value of the products for the period 1880–9.

2. From the enrollment in colleges, universities, and schools of technology of the United States for the following years:

1900	98,923
1905	120,099
1910	163,019
1915	237,011

Estimate the enrollment for the period 1905–9; the period 1910–14.

3. Same as problem (2) but using the following data:

1899	92,385
1904	111,688
1909	161,808
1914	216,636

4. Estimate the total consumption of sugar per capita in the United States for the period 1905–9, from the figures:

1900	58.81 pounds
1905	71.55
1910	79.77
1915	86.84

5. Estimate the consumption of sugar per capita (from data of problem (4)) for the period 1900–4; for the period 1901–5.

6. Estimate the consumption of sugar per capita for the period 1911–15.

EXERCISES[2] IN CONSTRUCTING ABRIDGED MORTALITY TABLES

1. The population statistics (decennial) of the whole country, and the mortality statistics (annual) of the so-called registration area, are published only by age groups, because of the concentration at ages which are multiples of 5 when the statistics are collected by single

[2] These Exercises are perhaps a little too specialized to be included in a short course.

ages. The population statistics (L_x) for females of ten registration states for 1920, for decennial age groups, are in part as follows:

Age x	L_x
5–14	2,529,809
15–24	2,350,086
25–34	2,348,953
etc.	

Interpolate the populations for the single ages 10 and 20.

Ans. 250,921 and 233,842.

2. The mortality statistics (d_x) for females of ten registration states for 1920 are in part as follows:

Age x	d_x
5–14	6,493
15–24	10,609
25–34	15,866
etc.	

Interpolate the number of deaths for the single ages 10 and 20.

Ans. 662 and 1080.

(**3**) Compute the leading term and differences of the population statistics of Exercise 1, and interpolate the populations for all of the single ages 15–24.

(**4**) Same as Exercise 3, but for the mortality statistics of Exercise 2.

5. As the population (L_x) for any single age, as obtained from the federal statistics, is regarded as referring to the population at the middle of the calendar year, the population at the beginning of the year is approximated by adding one-half of the deaths that take place during the year. The death rate at any age is computed then by the relation

$$q_x = \frac{d_x}{L_x + \frac{1}{2}d_x},$$

and the probability of living one year by the relation,

$$p_x = 1 - q_x = \frac{L_x - \frac{1}{2}d_x}{L_x + \frac{1}{2}d_x}.$$

The values of p_x were computed by the latter relation by logarithms,

from statistics of which those given in the preceding exercises are a part, as follows:

Age x	$\log p_x$
10	9.9974–10
20	9.9954–10
30	9.9931–10
etc.	

Since the probability of living n years

$$ {}_np_x = p_x p_{x+1} p_{x+2} \cdots p_{x+n-1}, $$

then

$$ \log {}_np_x = \log p_x + \log p_{x+1} + \cdots + \log p_{x+n-1}. $$

Find the logarithm of the probability of living ten years for ages 10 and 20. (Hint: Use formulas (24) and (22)).

Ans. 9.9821−10 and 9.9377−10.

6. If we assume any specified number of individuals to be living at an early age, say age 10, we need only multiply by successive values of ${}_np_x$, say for $n=10$ and for ages 10, 20, 30, etc., to obtain the number of survivors (l_x) at isolated ages, say 20, 30, etc. Such a table of survivors for all ages, say 10, 11, 12, etc., constitutes what is called a mortality table; and a table of survivors at isolated ages, such as 10, 20, etc., is called an abridged mortality table. Wide variation in the death rates for ages in the neighborhood of the age of birth practically necessitates the omission of ages much before age 10 in the construction of abridged mortality tables; moreover, if the number assumed to be living at the earliest age, called the radix, is taken to be no greater than 1000, much trouble will be avoided in trying to trace the last survivors at the highest ages, which would, in any case, have no appreciable effect upon the most important or earlier parts of the table. The abridged mortality table constructed from the statistics represented above is as follows:

Age x	l_x	Age x	l_x
10	1000	60	646
20	965	70	438
30	913	80	183
40	847	90	24
50	768		

Check the values of l_{20} and l_{30}.

7. Mortality tables are almost always accompanied by a column of values of what is called the expectation of life (e_x). The formula for computing the expectation of life at any age x is

$$e_x = \frac{1}{2} + \frac{l_{x+1} + l_{x+2} + \text{to end of table}}{l_x}.$$

The expression on the right, omitting the $\frac{1}{2}$, would give the expectation of life (called the curtate expectation of life) if all survivors died at the very beginning of the year of death; as the deaths are much more likely to be distributed uniformly throughout the year of death, it is usually assumed that the average person lives half a year in the year of death and so $\frac{1}{2}$ is added.

If formula (23) were applied to the abridged mortality table given above (formula (25) should be used to interpolate the first value), the values of l_x obtained for the age groups (11–20), 21–30, 31–40, etc., when added accumulatively, beginning with the higher ages, would give successive values which, when divided by the corresponding values of l_x for individual ages, would give values of the curtate expectation of life. Some method of determining the value of l_x for the age group 91–100 would have to devised, but no great concern should be felt in this determination (graphical methods would be satisfactory) because its value can have no appreciable effect upon the values of the expectation of life at the earlier and most important ages—the only ages at which the statistics are usually sufficiently reliable.

Check the following work:

Age	l_x	$l_{(x+1)-(x+10)}$	Σ	Curtate Expectation	e_x
10	1000	9821	52,395	52.4	52.9
20	965	9377	42,574	44.1	44.6
30	913	8779	33,197	36.4	36.9
40	847	8046	24,418	28.8	29.3
50	768	7044	16,372	21.3	21.8
60	646	5388	9,328	14.4	14.9
70	438	3017	3,940	9.0	9.5
80	183	875	923	5.0	5.5
90	24	48	48		

8. The population and the mortality statistics for 1920, of the males of the ten states which were registration states in 1900, are as follows:

Ages	Population	Deaths	Age	e_x
5–14	2,535,630	7,582	10	52.3
15–24	2,236,577	9,979	20	44.0
25–34	2,396,321	14,681	30	36.1
35–44	2,029,859	16,816	40	28.4
45–54	1,562,933	20,348	50	21.0
55–64	963,954	25,602	60	14.3
65–74	488,935	28,525	70	9.0
75–84	169,295	21,717	80	5.2
85–94	23,494	6,046		
95–	940	327		

Construct an abridged mortality table for decennial ages and check the column of values of expectation of life given to the right above.

CHAPTER IV

GAMMA AND BETA FUNCTIONS

29. The Gamma Function: Integrations by Substitution: Indeterminant Forms.—There are several types of definite integrals in the ordinary calculus which do not conform strictly or definitely to the definitions laid down for the ordinary types. Most of these special types are included under and referred to as *improper* integrals and refer graphically to areas which, though finite, stretch away to infinity in at least one direction. One of these types, known as the *Gamma function*, is unusually valuable in the mathematical theory of statistics, especially in the treatment of frequency curves; and even though we give no attention here to the systematic treatment of frequency curves, some knowledge of the function is highly desirable. Moreover, a certain development of the function will lead to an important relation between certain forms of the function and the definite integrals of the equation of a very important curve —the normal curve—to be considered later. It should be emphasized, however, that the treatment given here is necessarily elementary.

The Gamma function may be defined by the definite integral

$$\Gamma(n+1) = \int_0^\infty x^n e^{-x} dx. \quad . \quad . \quad . \quad . \quad . \quad (26)$$

Reference to any good textbook [1] on the calculus will reveal the fact that, although the area under consideration stretches away without limit to the right, the values of the

[1] See Byerly's "Integral Calculus," p. 86. Sometimes the Gaussian symbol $\pi(n)$ is used instead of the symbol $\Gamma(n+1)$ which is due to Legendre.

Gamma function are finite and determinate for all finite values of n greater than -1 (and infinite otherwise).

There are many definite integrals which bear no apparent resemblance to that given by (26) and which are, nevertheless, essentially Gamma functions, as suitable substitutions will show. The student is supposed to be familiar with the process of integration by substitution, but the process is so important, particularly in this chapter, that a typical example will be treated in detail. The process to be emphasized here is so definite that the student should have no trouble in handling similar integrations which follow. As an example, we shall substitute $x=y/a$ in the following integral to obtain the more general relation

$$\int_0^\infty x^n e^{-ax}dx=\frac{\Gamma(n+1)}{a^{n+1}}. \quad . \quad . \quad . \quad . \quad . \quad . \quad (27)$$

In making this substitution it is simply important that the substitution be made properly effective *in three places*: the integrand, the differential, and the limits.

The integrand evidently becomes $\frac{y^n}{a^n}e^{-y}$ and the differential dx becomes $\frac{dy}{a}$.

The limits on y corresponding to the limits 0 and ∞ on x, obtained by substituting these limits for x in the equation $y=ax$, are likewise 0 and ∞. Making these substitutions we have

$$\int_0^\infty x^n e^{-ax}dx=\frac{1}{a^{n+1}}\int_0^\infty y^n e^{-y}dy=\frac{\Gamma(n+1)}{a^{n+1}}.$$

Another important relation is obtained if we integrate (26) by parts (with $u=x^n$, etc.) to obtain

$$\int_0^\infty x^n e^{-x}dx=(-x^n e^{-x})_0^\infty+n\int_0^\infty x^{n-1}e^{-x}dx.$$

Here we need to refer to another important problem in the calculus—the valuation of indeterminate forms. It will be

recalled that expressions which take either of the forms $\frac{0}{0}$ or $\frac{\infty}{\infty}$ for some value of the independent variable can usually be valued by taking the derivative of the numerator for a new numerator and the derivative of the denominator for a new denominator, or by sufficient repetitions of the process. It will be seen that the first expression on the right of the equation given above takes the indeterminate form $\frac{\infty}{\infty}$ for $x=\infty$, and the application of the process just outlined shows that the limiting value of the expression is zero; its value for $x=0$ is obviously zero. We have then the important relation

$$\Gamma(n+1)=n\Gamma(n). \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (28)$$

If n is a positive integer

$$\Gamma(n+1)=n!.$$

Moreover, according to (26)

$$\Gamma(1)=\int_0^\infty e^{-x}dx=1$$

and it follows from (28) that

$$\Gamma(2)=1.$$

It is evident from formula (28) that if the value of the Gamma function is known for all positive numbers located between any two successive positive integers, say between 1 and 2, the value for any other positive number can be easily determined. For example, $\Gamma(3.36)=(2.36)(1.36)\Gamma(1.36)$. Excellent tables [2] of the logarithms of values of the Gamma function of numbers lying between 1 and 2 have been constructed. A small table follows.

[2] Pearson's "Tables" (7 places).
Legendre's Works, Vol. II (12 places).

A Short Table of Values of $10+\log \Gamma(n)$

n	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.0	9.9999	.9975	.9951	.9928	.9905	.9883	.9862	.9841	.9821	.9802
1.1	.9783	.9765	.9748	.9731	.9715	.9699	.9684	.9669	.9655	.9642
1.2	.9629	.9617	.9605	.9594	.9583	.9573	.9564	.9554	.9546	.9538
1.3	.9530	.9523	.9516	.9510	.9505	.9500	.9495	.9491	.9487	.9483
1.4	.9481	.9478	.9476	.9475	.9473	.9473	.9472	.9473	.9473	.9474
1.5	.9475	.9477	.9479	.9482	.9485	.9488	.9492	.9496	.9501	.9506
1.6	.9511	.9517	.9523	.9529	.9536	.9543	.9550	.9558	.9566	.9575
1.7	.9584	.9593	.9603	.9613	.9623	.9633	.9644	.9656	.9667	.9679
1.8	.9691	.9704	.9717	.9730	.9743	.9757	.9771	.9786	.9800	.9815
1.9	.9831	.9846	.9862	.9878	.9895	.9912	.9929	.9946	.9964	.9982

In using this table it should be noticed in interpolation that there is a minimum value of $10+\log \Gamma(n)$ in the neighborhood of $n=1.46$.

EXERCISES

Find the limiting values of the following:

1. $\dfrac{x^2-4}{x-2}$ for $x=2$. *Ans.* 4.

2. $\dfrac{x^2-16}{x^2+x-20}$ for $x=4$. *Ans.* $\frac{8}{9}$.

3. $\dfrac{3x^2-4x}{2x^2-3x+1}$ for $x=\infty$. *Ans.* $\frac{3}{2}$.

4. $\dfrac{\log_e x}{x^2}$ for $x=\infty$.

5. xe^{-x} for $x=\infty$.

6. $\dfrac{x}{2^x}$ for $x=\infty$.

7.[3] $x \log x$ for $x=0$. $\left(\text{Hint: write it } \dfrac{\log x}{1/x}.\right)$ *Ans.* 0.

[3] The expression given in this exercise takes the indeterminate form $0 \cdot \infty$, but this form and others, such as $\infty - \infty$, 1^∞ (see Ex. 9) etc., can usually be expressed to take either of the forms $\dfrac{0}{0}$ or $\dfrac{\infty}{\infty}$.

8. $x^2 \log x$ for $x=0$. $\left(\text{Write it } \frac{\log x}{x^{-2}}.\right)$ *Ans.* 0.

9. $1+i=\left(1+\frac{1}{x}\right)^{jx}$ for $x=\infty$.

$$\left(\text{Hint: } \frac{1}{j}\log(1+i)=x\log\left(1+\frac{1}{x}\right)=\frac{\log\left(1+\frac{1}{x}\right)}{1/x}.\right)$$

Ans. $\log(1+i)=j$,
or $1+i=e^j$.

30. The Beta Function.—Another definite integral which is very important in the mathematical theory of statistics, and which is closely related to the Gamma function, is the *Beta function*, or the First Eulerial Integral (the Gamma function is sometimes referred to as the Second Eulerial Integral), which may be defined by the relation

$$\beta(m, n)=\int_0^1 x^{m-1}(1-x)^{n-1}dx. \quad . \quad . \quad . \quad (29)$$

This integral can be shown [4] to be finite and determinate in value for all positive values of m and n.

If we substitute $1-y$ for x in this integral, in the manner outlined in the preceding section, we obtain the relation

$$\beta(m, n)=\beta(n, m), \quad . \quad . \quad . \quad . \quad . \quad . \quad (30)$$

which shows that the values of m and n are commutative.

If, moreover, we substitute $\frac{x}{1+x}$ and $\frac{1}{1+x}$ successively for x in (29) we obtain

$$\beta(m, n)=\int_0^\infty \frac{x^{m-1}}{(1+x)^{m+n}}dx \quad . \quad . \quad . \quad . \quad (31a)$$

$$=\int_0^\infty \frac{x^{n-1}}{(1+x)^{m+n}}dx, \quad . \quad . \quad . \quad . \quad (31b)$$

which are, therefore, merely other forms of the Beta function.

[4] Byerly's "Integral Calculus," p. 113.

The value of any Beta function for positive values of m and n can be expressed directly in terms of Gamma functions and can therefore be found by means of a table of values of the Gamma function. It happens that this relation between the Beta function and the Gamma function can be obtained by merely valuating a certain double integral in two ways and equating the results on the assumption that the order of integration (i.e., first with respect to x and then with respect to y, or first with respect to y, etc.) is immaterial. To value the integrals in this case with respect to either variable and in either order, one needs merely to concentrate his attention upon the essential factors and recognize the combination necessary at the time as constituting the Gamma function or the Beta function—as the case may be. The double integral is as follows:

$$\int_0^\infty \int_0^\infty e^{-x(y+1)} x^{m+n-1} y^{n-1} dx\, dy.$$

It is easily verified that the integral with respect to x (treating y-terms as constants) is

$$\Gamma(m+n) \int_0^\infty \frac{y^{n-1}}{(y+1)^{m+n}} dy = \Gamma(m+n)\beta(m, n). \quad . \quad . \quad . \quad (A)$$

If, however, we integrate first with respect to y we obtain

$$\Gamma(n) \int_0^\infty x^{m-1} e^{-x} dx = \Gamma(n)\Gamma(m). \quad . \quad . \quad . \quad . \quad . \quad (B)$$

Equating (A) and (B) we obtain

$$\beta(m, n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}. \quad . \quad . \quad . \quad . \quad . \quad (32)$$

If we let $m = n = \frac{1}{2}$ in $(31a)$ or $(31b)$ we obtain a form which we can easily integrate to give

$$\beta(\tfrac{1}{2}, \tfrac{1}{2}) = \pi.$$

But, according to (32),

$$\Gamma(\tfrac{1}{2}) = \sqrt{\beta(\tfrac{1}{2}, \tfrac{1}{2})}.$$

Hence,

$$\Gamma(\tfrac{1}{2}) = \int_0^\infty x^{-\frac{1}{2}} e^{-x} dx = \sqrt{\pi}.$$

It is easily verified that if we substitute x^2 for x in the last integral we obtain

$$\Gamma(\tfrac{1}{2}) = 2\int_0^\infty e^{-x^2} dx = \sqrt{\pi}.$$

Now $y = e^{-x^2}$ is the simplest form of the equation of the normal curve, which we shall consider at some length later, and the fact that x occurs only to the second degree shows that the graph is symmetrical with respect to the y-axis. Hence, twice the area under the curve from $x = 0$ to $x = \infty$ is the same as the area under the whole curve and we may write the relation given above as follows:

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}. \quad . \quad . \quad . \quad . \quad . \quad . \quad (33)$$

If, in addition, we substitute $x/\sqrt{h}$ for x in (33) we obtain the more general result

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{h}} dx = \sqrt{h\pi}, \quad . \quad . \quad . \quad . \quad . \quad (34)$$

which we shall find useful later, in connection with the normal curve.

EXERCISES

Many valuable exercises will be found in verifying the various relations given in the text.

1. Show that $\int_0^\infty x^{-\frac{1}{2}} e^{-x} dx = \sqrt{\pi}$.
2. Determine the value of π from the relation $\Gamma(0.5) = \sqrt{\pi}$.
3. Show that $\int_0^1 \left(\log \frac{1}{x}\right)^{n-1} dx = \Gamma(n)$.

   Substitute e^{-x} for x.

4. The Psi (ψ) function is defined as the derivative of the logarithm of the corresponding Gamma function. Given that

$$\Gamma(a+x)=(a+x-1)(a+x-2)\ \ldots\ x\Gamma(x),$$

where a is a positive integer, show that,

$$\psi(a+x)=\psi(x)+\frac{1}{x}+\frac{1}{x+1}+\ \ldots\ +\frac{1}{x+a-1}.$$

5. Show that $\beta(m, n)\beta(m+n, l)=\dfrac{\Gamma(l)\Gamma(m)\Gamma(n)}{\Gamma(l+m+n)}$.

6. Show that $\dfrac{\beta(l+m, n)}{\beta(m+n, l)}=\dfrac{\Gamma(l+m)\Gamma(n)}{\Gamma(m+n)\Gamma(l)}$.

7. Show that $\displaystyle\int_0^a x^{l-1}(a-x)^{m-1}dx=a^{l+m-1}\beta(l, m)$.

Substitute x/a for x.

8. Show that $\displaystyle\int_0^1 x^{l-1}(1-x^n)^{m-1}dx=\frac{1}{n}\beta\left(\frac{l}{n}, m\right)$.

Substitute x^n for x.

9. Show that $\displaystyle\int_0^1\frac{x^{m-1}+x^{n-1}}{(1+x)^{m+n}}dx=\beta(m, n)$.

Add together formulas (31a) and (31b), separate result into $\displaystyle\int_0^1+\int_1^\infty$ and substitute $1/x$ for x in the last part.

10. Show that the number of combinations of n things taken r at a time or

$${}_nC_r\left(=\frac{n!}{(n-r)!\,r!}\right)=\frac{1}{\beta(n-r, r)}.$$

11. Show that $\beta(6, 4)=1/210$.

12. Show that $\beta(1, 1)=1$.

13. Find the values of (a) $\log\beta(2.36, 2.49)$.

(b) $\log\beta(3.4, 3.6)$.

CHAPTER V

PROBABILITY

31. A Priori Probability: Simple Events.—If a bag contains three white and five black balls, and one ball is drawn out at random, what is the probability that this ball is white? The event in question is said to happen if a white ball is drawn and to fail if a black ball is drawn. The number of distinct ways in which the event may happen is three and the total number of possible ways in which it may happen or fail is eight. The fraction $\frac{3}{8}$ then is said to be the probability of drawing a white ball. This illustrates the following definition of probability:

If all the happenings and failings of an event can be analyzed into $h+f$ possible ways, each of which is equally likely, and if in h of these ways the event will happen and in f of them fail, the probability that the event will happen is $\frac{h}{h+f}$ *and the probability that it will fail is* $\frac{f}{h+f}$.

The *chances* of the event happening are said to be as h is to f.

Corollary. *The sum of the probability that an event will happen and the probability that it will fail is* 1, *which is the symbol for certainty.* The symbol for certain failure is 0.

In applying the definition of probability, the fact should not be overlooked that all the ways are assumed to be equally likely. To illustrate the need of precaution in this matter, consider the question: What is the probability of throwing "head" at least once in two throws of a coin? We might give the following as the equally likely cases: HH, HT and TT; whence, the probability would be $\frac{2}{3}$. Further consideration,

however, makes it clear that the case HT is twice as likely as HH or TT because it can happen in two ways, that is, HT or TH. The probability desired is therefore $\frac{2}{4}$.

The probabilities referred to above are said to be determined *a priori* and are often referred to as *a priori* probabilities to distinguish them from probabilities which will be considered later and which can not be determined by an *a priori* analysis of the various possibilities.

We shall refer to events which can only happen or fail as *simple* events to distinguish them from events which involve other possibilities.

Sometimes the formula for the number of different combinations of n things taken r at a time proves useful in determining an *a priori* probability, especially when the total number of ways an event can happen is a large number. This formula is

$$ {}_nC_r = \frac{n(n-1)\ \ldots\ (n-r+1)}{r!} \quad \text{or} \quad \frac{n!}{r!(n-r)!}. $$

As an example, let us determine the probability of drawing two white balls and three black balls in drawing five balls at random from a bag containing four white balls and six black balls. The total number of different ways of drawing five balls from a bag containing ten balls is ${}_{10}C_5 = \frac{10\cdot 9\cdot 8\cdot 7\cdot 6}{5\cdot 4\cdot 3\cdot 2} = 252$. The number of different ways of drawing two white balls from four white balls is ${}_4C_2 = \frac{4\cdot 3}{2} = 6$, and the number of ways of drawing three black balls from six black balls is ${}_6C_3 = 20$; hence, the number of ways of drawing two white balls and three black balls is $6\cdot 20 = 120$. The desired probability is then $120/252 = 10/21$.

EXERCISES

1. A bag contains 3 red, 4 black and 5 white balls; if 1 ball is drawn at random, what is the probability that it is a red ball? A red or a black ball?

2. Three dice are thrown.

(a) What is the probability of throwing a 7 (the sum of the upper faces)?

(b) Show that the probability of throwing a 14 is five times that of throwing a 4.

(c) Show that the probabilities of throwing a 10 or an 11 are the same. What is the probability?

3. If the probability of a certain event happening is 4 times the probability of its failing, what is the probability of its happening?

4. What is the probability of throwing a head in 3 throws of a coin?

5. What is the probability of throwing an ace in 6 throws of a die?

6. What is the probability of throwing exactly 1 head in 3 throws of a coin?

7. A bag contains 4 white and 6 black balls; find the probability of drawing exactly 2 white balls out of 5 drawn at random. At least 2 white balls.

8. A purse contains 2 dimes, 3 quarters and 4 half-dollars. Assuming that one coin is as likely to be drawn as another, what is the probability that if a single coin is drawn it will be either a quarter or a half-dollar?

9. If 12 students are seated at random in a row, what is the probability that A and B are next to each other?

10. Three balls are drawn at random from a bag containing 5 black and 4 white balls. What is the probability that 2 are black and 1 white?

11. Two cards are drawn at random from a suit of 13 cards. What is the probability that the 2 cards are an ace and a king?

32. Independent Events.—Two or more events are said to be *independent* when the occurrence of any one of them is not affected by the occurrence or non-occurrence of any of the rest. Thus, the results of two drawings of a ball from a bag are independent if the ball is returned after the first drawing, but interdependent if the ball is not returned.

Theorem.—*The probability that all of a set of independent events will occur is the product of the probabilities of the single events.*

For, consider two such events whose probabilities are a/n and b/m respectively. The number of equally likely possible cases for and against the first event is n, for and against the second m, and since the events are independent any one of the n cases may occur with any one of the m cases. Hence, the number of equally likely cases for and against the occurrence of both events is nm. By the same reasoning, ab of these cases favor the occurrence of both events. Therefore, the probability that both events will occur is ab/nm or $\frac{a}{n}\cdot\frac{b}{m}$. The demonstration for the case of more than two events is similar.

Thus, the probability of throwing an ace twice in succession with a single die is $\frac{1}{6}\cdot\frac{1}{6}=\frac{1}{36}$. Likewise, the chance of drawing a white ball twice in succession from a bag which contains four white and three black balls, the ball first drawn being returned before the second drawing, is $\frac{4}{7}\cdot\frac{4}{7}=\frac{16}{49}$.

33. Mutually Exclusive Events.—If two or more events are so related that but one of them can occur, they are said to be *mutually exclusive.* Thus, the throwing of a "head" and the throwing of a "tail" in the same throw of a single coin are mutually exclusive events.

Theorem.—*The probability that some one or other of a set of mutually exclusive events will occur is the sum of the probabilities of the single events.*

For, consider two mutually exclusive events A and B. The possible cases with respect to the two events are of three kinds, all mutually exclusive, namely, those for which (1) A happens, B fails; (2) A fails, B happens; (3) A fails, B fails. Let the numbers of equally possible cases of these three kinds be l, m and n respectively. Then the probability of the single event A is

$$\frac{l}{l+(m+n)},$$

for, since A never happens except when B fails, the l cases in which A happens and B fails are *all* the cases in which A happens, and the $m+n$ cases in which A fails and B happens or A and B both fail are all the cases in which A fails.

Similarly, the probability of the single event B is

$$\frac{m}{m+(l+n)}.$$

Hence, the probability that either A or B happens or

$$\frac{l+m}{l+m+n}=\frac{l}{l+(m+n)}+\frac{m}{m+(l+n)}.$$

The proof for more than two events is similar.

Thus, if one ball be drawn from a bag containing 3 white, 5 black and 7 red balls, since the probability of its being white is $\frac{1}{5}$ and that of its being black is $\frac{1}{3}$, the probability of its being either white or black is $\frac{1}{5}+\frac{1}{3}=\frac{8}{15}$.

Care must be taken to apply this theorem only when events are mutually exclusive. Thus, if asked to find the probability that a problem will be solved if both A and B attempt it, A's probability of success being $\frac{3}{4}$ and B's $\frac{2}{3}$, we can not obtain the desired result by merely adding $\frac{3}{4}$ and $\frac{2}{3}$, since the two events (A succeeds and B succeeds) are not mutually exclusive. The mutually exclusive cases are: A succeeds, B fails; A fails, B succeeds; A succeeds, B succeeds. The probabilities of these events are $\frac{3}{12}$ $(=\frac{3}{4}\cdot\frac{1}{3})$, $\frac{2}{12}$ $(=\frac{1}{4}\cdot\frac{2}{3})$ and $\frac{6}{12}$ $(=\frac{3}{4}\cdot\frac{2}{3})$ respectively; and the sum of these probabilities or $\frac{11}{12}$ is the probability that the problem will be solved. This problem could also be solved as follows: the probability that both will fail is $\frac{1}{4}\cdot\frac{1}{3}=\frac{1}{12}$; the probability that both will not fail—that is, that at least one will solve the problem—is then $1-\frac{1}{12}$, or $\frac{11}{12}$.

EXERCISES

1. If the probability that A will live ten years is $\frac{7}{8}$ and the probability that B will live ten years is $\frac{9}{10}$, what is the probability that

both will be alive after ten years? What is the probability that one or the other will be alive then?

2. A bag contains 2 white, 3 black and 4 red balls. What is the probability that a ball drawn at random will be either white or red?

3. What is the probability of throwing either an ace or a deuce in a throw of 2 dice?

4. A traveler has five connections to make in order that he may reach his final destination on time. If his estimates that for each of these connections the chances are 2 to 1 in his favor are correct, what is the probability of his making all his connections?

5. If the probability that A and B will survive a certain period is $\frac{7}{8}$ and $\frac{9}{10}$ respectively, what is the probability that:

(*a*) one or the other will die in the period?
(*b*) exactly one will die in the period?
(*c*) both will die in the period?

6. If the probability that each of n individuals will survive a certain period is p, what is the probability that at least one will die in the period?

7. If the probability that the age of a man selected at random from a group of men is between 20 and 25 years is $\frac{1}{4}$, and the probability that it is between 25 and 35 is $\frac{1}{3}$, what is the probability that his age is between 20 and 35?

8. The probability that A will solve a problem if he attempts it is $\frac{1}{5}$, and that of B $\frac{1}{6}$. What is the probability that the problem will be solved if both try it? What is the probability that exactly one of them will solve it? What is the probability that both will solve it?

9. A bag contains 3 red, 4 black and 5 white balls. Suppose that 2 balls are drawn at random; what is the probability that:

(*a*) both are red?
(*b*) both are black?
(*c*) both are either red or (both) black?
(*d*) they consist of exactly one red and one black?
(*e*) they are either black or red?
(*f*) both are white?

(*g*) that exactly one is white?
(*h*) at least one is white?
(*i*) Solve (*e*) using results of (*f*) and (*g*).

10. If the probabilities of A, B, C and D surviving a certain period are $\frac{6}{7}$, $\frac{7}{8}$, $\frac{8}{9}$ and $\frac{9}{10}$ respectively, what is the probability that at least one of the four will die in the period?

11. If the probabilities of A, B, C and D dying in a certain period are $\frac{1}{7}$, $\frac{1}{8}$, $\frac{1}{9}$ and $\frac{1}{10}$ respectively, what is the probability that at least one of the four will die in the period? What is the probability that all will survive the period?

34. Empirical Probability: Homogeneous Populations.— The fraction $\frac{h}{h+f}$, which we have called the *a priori* probability of an event, means little so far as the actual outcome of a single trial or a small number of trials of the event is concerned. It should, however, indicate the frequency with which the event would occur in the long run, that is, in the course of an indefinitely long series of trials. Thus, if one should try the experiment of throwing a coin a very great number of times, say several thousand times, one would find that, as the number of throws increases, the ratio of the number of times that a head appears to the total number of throws approaches the value $\frac{1}{2}$ more and more closely and steadily. In general, if h be used to refer to the number of times a certain event occurs and n refers to the number of trials, and p the corresponding probability, then the value of this probability may be defined by the relation

$$p = \lim_{n \to \infty} \frac{h}{n}. \quad \ldots\ldots \quad (35)$$

Or the value of the probability in question may be defined as the limiting value of the ratio h/n as n, the number of trials, increases without limit. Such a probability is called an *empirical* or *a posteriori* probability.

Now, it is only proper that we approach the subject of

probabilities from the point of view of *a priori* probabilities, or probabilities that are associated with games of chance; but the student will soon observe that the vast majority of the most important problems in determining the values of probabilities in practice require an entirely different mode of procedure. Thus, if we wish to determine the probability that a man aged 30 will die in the following year, we have no way of analyzing all the happenings and failings which are equally likely; in fact, we know that the chances of a number of men aged 30 dying in the following year are not at all equal. We have noticed, however, that whatever the value of the desired probability may be, it will *tend* to be exemplified as the number of trials increases indefinitely. This fact suggests that such a probability be determined empirically.

The author wishes now to call attention to a very important distinction. The total field of events, or *population*, of a given investigation may be far from *homogeneous*; and since it is obviously impossible to consider all events in determining an empirical probability whose population is infinite, we are left with the sole possibility of taking what we shall call a *random sample* of the whole population, hoping to arrive at an approximation of the desired probability. If the discrepancy proves to be no greater relatively than that which we usually find in properly conducted games of chance, the sample is said to be *representative* of the whole population. If the discrepancy is obviously excessive, we conclude either that the sample was not selected at random or that the population is not sufficiently homogeneous. We shall find, as we proceed, that we can control the discrepancies or errors due to random sampling in a homogeneous population; but we can not always be sure that we have a random sample, and it is particularly difficult to insure that we are dealing with a population which is sufficiently homogeneous. As it is characteristic of the great majority of empirical investigations in various fields that the results must be obtained from a sample, the question of the homogeneity of the population is exceedingly important. In

the absence of definite knowledge of the homogeneity of a population, the wise investigator will regard new results merely as tentative and subject to check by results of similar character.

A method of investigating the homogeneity of a given population will be considered in a later chapter, but it will be necessary to assume for the present that the populations considered in the immediately succeeding chapters are homogeneous.

EXERCISES IN SURVIVORSHIP

1. We shall denote the probability that a person aged x will survive n years by ${}_np_x$ and the product ${}_np_x \cdot {}_np_y$ by ${}_np_{xy}$. Given, then, that two persons are aged x and y, express the probability that:

(a) both will live n years;
(b) both will die within n years;
(c) both will not live n years;
(d) both will not die within n years;
(e) at least one will live n years;
(f) at least one will die within n years;
(g) exactly one will live n years;
(h) exactly one will die within n years.

2. Given three persons aged x, y and z, express the probability that:

(a) all three will live n years;
(b) all three will not live n years;
(c) at least one will die within n years;
(d) at least one will live n years;
(e) at least two will live n years;
(f) at least two will die within n years;
(g) all three will not die within n years;
(h) exactly two will live n years;
(i) exactly two will die within n years;
(j) no more than one will live n years;
(k) no more than one will die within n years.

3. To what events do the following probabilities refer?

(*a*) $1-{}_np_{xy}$;

(*b*) $1-{}_np_{xyz}$;

(*c*) $(1-{}_np_x)(1-{}_np_y)$;

(*d*) $1-(1-{}_np_x)(1-{}_np_y)(1-{}_np_z)$;

(*e*) ${}_np_x+{}_np_y$ (criticize);

(*f*) ${}_{n-1}p_x-{}_np_x$;

(*g*) ${}_np_x(1-{}_np_y)+{}_np_y(1-{}_np_x)$;

(*h*) ${}_np_x-{}_np_{xy}$;

(*i*) ${}_np_x\cdot{}_np_{xy}$ (criticize);

(*j*) ${}_np_x+{}_np_y-{}_np_{xy}$;

(*k*) ${}_np_x+{}_np_y-2{}_np_{xy}$.

35. Repeated Trials of a Single Event.—The following theorems are concerned with the question of the chance that a certain event will occur a specified number of times in the course of a series of trials, the chance of its occurrence in a single trial being known.

Theorem.—*If the probability that an event will occur in a single trial is p, the probability that it will occur exactly r times in the course of n trials is ${}_nC_rp^rq^{n-r}$, where $q=1-p$.*

For, the probability that it will occur in all of any particular set of r trials and fail on the remaining $n-r$ trials is p^rq^{n-r}. But since there are n trials all told, we may select this particular set of r trials in ${}_nC_r$ ways, which are, of course, mutually exclusive. Hence, the probability in question is ${}_nC_rp^rq^{n-r}$. Are all the events of such a particular set of r trials independent? Is this necessary? Explain. Why must the various sets of r trials be mutually exclusive?

Thus, the chance that ace will turn up exactly twice in 5 throws with a single die, or that out of 5 dice thrown simultaneously exactly 2 will turn up is ${}_5C_2(\frac{1}{6})^2(\frac{5}{6})^3$ or $\frac{6}{3}\frac{2}{8}\frac{5}{8}\frac{5}{8}$.

Observe that ${}_nC_rp^rq^{n-r}$ is the term containing p^r in the expansion of $(p+q)^n$. What is the number of the term?

Theorem.—*The probability that such an event will occur at least r times in the course of n trials is the sum of the first* $n-r+1$ *terms in the expansion of* $(p+q)^n$, namely

$$p^n+{}_nC_1p^{n-1}q+{}_nC_2p^{n-2}q^2+\ \ldots\ +{}_nC_rp^rq^{n-r}.$$

For, the event will occur *at least r* times if it occurs exactly r times or exactly any number of times greater than r, and the terms of the expansion given above refer to mutually exclusive events and represent the probabilities of the occurrence of the event exactly n, exactly $n-1$, . . . exactly r times, respectively. Thus, the chance that ace will turn up at least 4 times in the course of 5 throws with a single die is $(\frac{1}{6})^5+5(\frac{1}{6})^4(\frac{5}{6})=\frac{13}{3888}$.

EXERCISES

1. What is the probability that, in 6 throws of a coin, at least 3 will be heads?

2. Five coins are thrown. What is the probability that exactly 2 of them are heads? *Ans.* $\frac{10}{32}$.

3. Find the probability of throwing at least 8 in a single throw with 2 dice.

4. If A's probability of winning any single game against B is $\frac{3}{4}$, find the probability of his winning at least 3 games out of 7.

5. Show that the probability of throwing exactly r heads and $n-r$ tails in a single throw of n coins is ${}_nC_r\div 2^n$.

36. Cogent Reason and Insufficient Reason.—The author wishes now to call attention to a distinction between two principles of logic which are often confused in determining the values of probabilities. The distinction is very subtle and can not be established sufficiently rigorously to insure that one can recognize it clearly in all situations. The best that can be done here is to illustrate. We say that the probability of throwing a " head " with a single coin is $\frac{1}{2}$, because we believe that the throw is just as apt to prove " head " as " tail " or " tail "

as "head." But suppose that we have no more reason to believe that a certain book contains pictures than that it does not contain them; are we justified in assuming that the probability that the book contains pictures is $\frac{1}{2}$? The value of the former probability is said to be determined by *cogent reason* and the value of the latter probability—if it were acceptable—by *insufficient reason.* It should be obvious that, however subtle the situation may be, there is a little positive information in examples like the former, which removes them entirely from the class of examples like the latter, about which we know absolutely nothing. We simply say that examples of the latter type lie outside the domain of probabilities.

It is usually easy to establish the absurdity of a particular result of insufficient reasoning. To revert to the example considered above, we would be just as much justified in saying that the probability that the book contains pictures *of a certain type* is $\frac{1}{2}$ and, hence, that the probability that the book does not contain pictures of that type would also be $\frac{1}{2}$. Similarly, the probability that the book does not contain pictures of a second type would also be $\frac{1}{2}$, and so on. The probability, then, that the book does not contain pictures of any of a large number of types would be a large power of $\frac{1}{2}$ or a very small fraction. Finally then, the probability that the book contains pictures of at least one type would be unity minus this very small fraction, or approximately unity—a value quite different from $\frac{1}{2}$ which we assigned to the probability originally—and the contradiction is obvious.

The distinction made above is of great fundamental importance in the analysis of the observations in many investigations. As a simple example, if we were particularly anxious to obtain a very accurate reading of a finely graduated barometer, we should probably make a large number of readings and select the most probable value by methods to be introduced later, which would depend fundamentally upon cogent reason or belief that deviations to one side of the correct value are just as likely as deviations to the other side.

SOME FAMOUS PROBLEMS AND FALLACIES

1. D'Alembert believed the probability of throwing heads at least once in 2 throws of a coin to be $\frac{2}{3}$. Criticize.

2. Leibnitz thought that a throw of 12 with 2 dice was as probable as a throw of 11. Criticize.

3. Would the fact that a coin has been thrown to give heads several times in succession affect the probability of obtaining heads in the next throw? D'Alembert believed that it would. Moreover, Bequelin believed that if heads had been thrown n times in succession the probability that the next throw would yield a head would be $\frac{1}{n+1}$.

4. In seeking the probability of obtaining either 3 heads or 3 tails in a single throw of 3 coins, it has been reasoned that of 3 coins at least 2 must show heads or tails, and the probability that the third coin will be the same as the other 2 is $\frac{1}{2}$, and that the desired probability is therefore $\frac{1}{2}$. Locate the fallacy. What is the probability?

5. The famous problem known as the "martingale" consists in determining the relative chances of a poor man and a rich man who engage in a game of chance, in which the poor man continues, until he loses, to stake all he has against a like amount with the even chance of losing all or doubling his fund. The game ends, of course, whenever the poor man once loses. Cardan is said to have shown that the condition of play imposes a great disadvantage on the rich man, but we have no traces of his method of reasoning. What is your opinion of the relative chances of the two players? What does the rich man gain when the poor man once loses?

6. The famous St. Petersburg problem may be stated essentially as follows: A coin is tossed until head is obtained. Peter is to pay Paul 1 dollar if head appears for the first time on the first toss, 2 dollars if on the second toss, and, in general, 2^{n-1} dollars if on the n-th toss. What is Paul's expectation,[1] or what should Paul pay Peter at the outset so that the play will be fair to both? Show that

[1] Mathematical expectation of an event is defined as the product of the probability of the occurrence of the event and the amount to be gained if the event occurs.

if a sufficient number of games were played Paul would stand to gain any amount, however large.

7. Assuming that a certain gambler will win 3 games out of 5 in which he plays, and in which he always stakes $\frac{1}{2}$ of his funds against an equal amount, show that he must inevitably lose, by showing that his fund after playing $5n$ games would be $(\frac{2}{3}\frac{7}{2})^n$ times his original fund.

MISCELLANEOUS EXERCISES

1. An illiterate servant places 12 similar books on a shelf. What is the probability that 3 volumes of a set are together? What is the probability that they are together in their proper order?

2. Show that the probability of a leap year containing 53 Sundays is $\frac{2}{7}$.

3. If 4 cards are drawn from a pack of 52 cards, show that the probability of there being 1 from each of the suits is $\frac{13^3}{49 \cdot 25 \cdot 17}$.

4. Two of 8 keys on a ring will open a certain door. What is the probability that the door can be unlocked by 1 of 3 keys selected at random?

5. Each student in a certain class of 10 is likely to make a sufficiently correct observation of the sun's transit 1 time out of 4. What is the probability that a given transit will be correctly observed by the entire class? What is the probability that it will be correctly observed at all if the entire class attempts it?

6. A party of 10 attending the opera are to occupy 6 seats in one group and 4 seats in another group. If the division is made at random, what is the probability of 2 given individuals, A and B, being in the same group?

7. What is the probability of a player in a game of whist holding 3 aces? At least 3 aces?

8. One purse contains 6 silver dollars, and 4 quarters. Another contains 2 silver dollars and 10 quarters. If a purse is selected at random and a coin extracted at random, what is the probability that the coin selected will be a dollar?

9. If 5 books are brought at random from a shelf of 20, what is the probability that 3 desired books will be among them?

10. What is the probability of drawing 4 red balls, 2 white balls and 5 black balls, in drawing 11 balls at random from a bag containing 7 red, 6 white and 9 black balls?

11. What is the probability of throwing an ace in 3 throws of a die?

12. What is the probability of throwing a head in n throws of a coin?

13. What is the probability of throwing an ace exactly once in 3 throws of a die?

14. What is the probability of throwing a 10 (the sum of the upper faces) with 3 dice? Of throwing a 9?

15. A bag contains 10 times as many white balls as black balls, and 1 ball is drawn at random. What is the probability that the ball drawn is white?

16. What is the expectation of a gambler who is to win \$30 if he throws a 17 with 3 dice?

17. A committee of 4 is to be selected at random from a group of 3 sophomores, 4 juniors and 5 seniors. What is the probability that the committee will consist of:

(*a*) 2 juniors and 2 seniors?
(*b*) 1 sophomore, 1 junior and 2 seniors?
(*c*) 4 seniors?

18. If 3 dice are thrown, what are the probabilities of throwing:

(*a*) 3 sixes?
(*b*) 2 sixes and a five?
(*c*) a six, a five and a four?

19. Given that it is an even chance that a certain ship will encounter a storm, the probability that the ship will spring a leak in the storm is $\frac{1}{10}$; if a leak occurs, the chances are 9 to 10 that the engine will pump her out, and if they fail the chances are 3 to 4 that the compartments will keep the ship afloat; and finally, if she sinks the chances are even that a traveler will be saved. What is the probability that the traveler will be lost at sea?

CHAPTER VI

AVERAGES AND AIDS IN THEIR COMPUTATION

37. Arithmetic Average, or Mean: The Geometric Average: The Median.—One of the main purposes included under the general heading of " analysis of statistics " is to set up a systematic method of computing the values of certain terms which will serve to describe a sample of numerical observations so well that a significant difference between the values of such a term corresponding to two samples will permit one to differentiate between the two fundamental situations. It should be obvious that if the value of such a term were sufficiently descriptive its quotation might well obviate the necessity of exhibiting the entire set of individual observations. The importance of the latter statement should be evident when it is pointed out that, in the ideal investigation, the number of observations will be made as large as possible, because such a procedure probably constitutes the best scheme to be followed in practice to insure that the value of the term under consideration shall be representative.

No one term is at the same time more useful and more familiar to the ordinary mind than the *arithmetic average*, or *mean*, of a set of numerical measurements or observations, which may be defined as *the sum of the values of the observations divided by their number*. As an example, it is easily verified that the arithmetic average of the following observations is $\frac{157.7}{36}$, or 4.38.

2.76	3.76	4.16	4.40	4.72	5.00
3.32	3.80	4.16	4.40	4.72	5.00
3.68	3.92	4.16	4.44	4.76	5.08
3.72	3.92	4.28	4.60	4.88	5.28
3.72	4.08	4.36	4.64	4.96	5.40
3.72	4.12	4.40	4.68	5.00	5.72

A statement of the meaning, etc., of the observations given above—which is immaterial for the present purpose—will be made later in another section.

It will be noticed that several of the values given above are the same. It should be evident that the same value would be found for the mean if, instead of summing the individual values, each value were first multiplied by the number of times it occurs and all such products were summed. Such an average is called, for purposes of distinction, the *weighted* arithmetic average. The set of observations given above would scarcely justify such a procedure, but we shall find that a vast majority of the arithmetic averages computed in practice will be essentially weighted arithmetic averages. Suppose that a set of observations were to occur with frequencies as indicated below:

Observations	Frequencies	
4.16	6	$4.16 \times 6 = 24.96$
4.20	7	29.40
4.40	10	44.00
4.46	9	40.14
4.52	4	18.08
	36	$36)\overline{156.58}$
		4.35

It would be only natural to compute the *weighted* arithmetic average in such a case. The natural method of computation is shown to the right. It should be noted that the value obtained (4.35) would be the same if it were computed strictly in accordance with the original definition of an arithmetic average.

Sometimes observations are *weighted* in exactly the same way as that illustrated above but for a different reason. For example, the five different observations suggested above might have been obtained in such a way that they would not be equally accurate, and there might be some reason why we should like to *weight* them in accordance with, say, the numbers which indicate frequencies above. Such a scheme is very common. For example, the various grades of a student—quizzes, recita-

tion, final examination, etc.—are almost invariably weighted in some way or other.

The term "average" is used quite widely to refer to the arithmetic average, and we shall follow that custom fairly consistently in the future, but there are other kinds of averages which would be preferable to the arithmetic average in some situations. The *geometric* average of a set of observations is obtained by multiplying the observations together and extracting the root corresponding to the number of the observations. For example, suppose that the population of a certain community gains by $\frac{1}{6}$, $\frac{1}{3}$ and $\frac{2}{3}$ of its population in three successive years, respectively. The average (geometric) annual rate of gain would be

$$\sqrt[3]{\tfrac{1}{6}\times\tfrac{1}{3}\times\tfrac{2}{3}}=\tfrac{1}{3}.$$

Sometimes it is desirable to weight the observations in determining the geometric average by raising each observation to its respective weight as a power.

The *median* of a set of observations is the middle observation when all the observations are *ranked* or arranged in order of magnitude. The median has the advantage, as an average, in that it requires no computation. It is easily found by inspection that the median of the set of observations given at the beginning of this section is 4.40. It will be noticed that, strictly speaking, there is no middle term in this case since there is an even number of observations; but since the value on each side of the middle is 4.40 we naturally select that value for the median. The median divides the number of observations into two equal or nearly equal parts. Similarly, *quartiles* divide the number of observations into four parts, *percentiles* into one hundred parts, etc.

Occasionally the *harmonic mean* is used, which is defined as the reciprocal of the arithmetic average of the reciprocals of the observations.

Certain other forms of averages will be mentioned in later sections. Since we shall give most of our attention to the

arithmetic average and certain other forms which are closely related to it, it should perhaps be stated here, once for all, that many cases arise in practice where other averages would be greatly preferred for one reason or another. We shall confine most of our attention to certain averages, such as the arithmetic average, mainly because of their possible mathematical content.

EXERCISES

1. According to Vilmorin's tables, the following seeds live approximately the designated numbers of years:

Bean	3	Egg Plant	6	Peas	3
Beet	6	Endive	10	Peanut	1
Broccoli	5	Kohlrabi	5	Pepper	4
Cabbage	5	Leek	3	Pumpkin	4
Carrot	4	Lentil	4	Radish	5
Cauliflower	5	Lettuce	5	Rape	5
Celery	8	Muskmelon	5	Salsify	2
Chicory	8	Nasturtium	5	Spinach	5
Corn	2	Okra	5	Squash	6
Corn Salad	5	Onion	2	Tomato	4
Cress	5	Parsnip	2	Turnip	5
Cucumber	10	Parsley	3	Watermelon	6

Find the average length of life of the seeds.

2. Ernest Thompson Seton gives, in "The Arctic Prairies," the number of antelopes in 26 bands seen along the C. P. Railroad in Alberta, within a stretch of 70 miles, as follows: 8, 4, 7, 18, 3, 9, 14, 1, 6, 12, 2, 8, 10, 1, 3, 4, 6, 18, 4, 25, 4, 34, 6, 5, 16, 4. Find the average number in a band.

3. The distribution of ages of pupils in a certain public school was as follows:

Ages	Frequencies
12	7
13	45
14	186
15	114
16	61
17	8

Find the average age.

4. A certain college gives a credit of 4, 3, 2, 1 points for each A, B, C, D respectively, received in a course. The average credits received by seniors above 3.1 in one semester were as follows:

Credit	4.0	3.9	3.8	3.7	3.6	3.5	3.4	3.3	3.2
Frequencies	11	1	8	8	5	13	11	11	19

What was the average credit received by this group?

5. The attendance at a certain university increased 10, 15 and 18 per cent in three successive years, respectively. What was the average (geometric) annual rate of increase?

6. If the protein content of corn of a community increased 5, 8, 12, 18 and 23 per cent in four successive years of breeding, what was the average (geometric) annual rate of increase?

7. Find the values of the medians in Exercises 1–4.

38. Frequency Distributions.—Attention was called in the preceding section to the possibility of two or more of a set of observations having the same value. It should be obvious that this possibility is due fundamentally to inaccuracy in expressing the observations and that if the measurements were sufficiently refined the expressed values of no two of them need be equal. Even then, however, a close examination would reveal the fact that the observations are by no means equally spaced but tend to concentrate at certain very important points. These points will be noted much more quickly and reliably if we sacrifice accuracy in expression to the extent of allowing the observations to fall into *classes* of equal intervals. The number of observations falling into any class will be called the *frequency*; the middle value of the possible measurements or observations of a class is called the *class mark*; and the complete series of pairs of class marks and corresponding frequencies, arranged in order of size of the class marks, is called a *frequency distribution*. The limiting measurements of a class are called the *class limits*. As an example, the 36 observations given in the preceding section may be classified to give any one of several possible frequency distributions; two of these possibilities are as follows:

No. 1		No. 2	
Observations (class marks)	Frequencies	Observations (class marks)	Frequencies
2.8	1	3.0	1
3.1	0	3.5	5
3.4	1	4.0	9
3.7	6	4.5	11
4.0	4	5.0	7
4.3	9	5.5	3
4.6	5		—
4.9	6		36
5.2	2		
5.5	1		
5.8	1		
	—		
	36		

The proper interpretation of the various items is very important. Thus, the frequency "9" in the second distribution refers to 9 observations not all of size 4.0 but of sizes ranging between the two limits 3.75 and 4.25. The frequency "9" in the first distribution refers to observations of sizes ranging from 4.15 to 4.45. The *class interval* of the second distribution is then 0.5 and that of the first is 0.3.

It is easily verified that the values of the arithmetic average of these distributions differ very little whether we use the isolated values as given originally or whether we use one of the frequency distributions given above. This fact holds fairly consistently for all distributions. We shall see later, moreover, that the labor of computing the value of the arithmetic average and of other expressions is greatly reduced by the use of frequency distributions. The formation and use of frequency distributions is further justified by the fact that in the ideal investigation, where a very large number of observations are made, the values of these observations fall naturally into classes from the very beginning.

It is only natural to ask how far one is justified in carrying the process of compressing a set of observations into a frequency

distribution. It should be emphasized that there may be important reasons for retaining such observations in their original form in some cases, but otherwise the criterion usually applied is that the compression of the observations may be carried until a relatively smooth series of frequencies is obtained. Such a criterion affords no rigid method of application, and we simply claim that the process of forming frequency distributions is natural and—as we shall see—leads to a much simpler method of computing the mean and other expressions with only a slight loss in accuracy.

39. Frequency Curves: Fitting Curves to Frequencies: The Mode.—If we plot the corresponding pairs of values of a frequency distribution, letting x refer to the observation or class mark and y to the corresponding frequency, and join the points thus obtained by a smooth curve, we obtain what is called a *frequency curve*. In many cases where the number of observations is small and where the discrepancies in the smoothness of the values of the frequencies are clearly due to this lack in number, it may be preferable to pass a smooth curve through as many points and as near the others as possible, in the hope of obtaining a representative curve which we shall refer to also as a frequency curve. The process of determining this representative curve, or of *fitting a curve to the observed frequencies*, is useful in smoothing or *graduating* these frequencies. Thus, if this representative curve were plotted carefully on ruled paper, the ordinates corresponding to the original observations could be easily read off to give a new and smoother distribution of frequencies, which is called a *graduation* of the original frequencies.

Other, and usually more satisfactory, methods of graduation will be considered later, which will consist essentially of determining the analytic expression for the frequency curve which fits the observed frequencies. Such a procedure will not only usually lead to more satisfactory graduations but, what is more important, will also render available the many processes of mathematical analysis. The fitting of curves to observed

frequencies by the formal processes of analysis is one of the most useful problems of scientific work, because it usually replaces a large number of relatively intangible statistical data by a single representative expression of algebraic form which may be analyzed at great length.

If a number of observations were arranged in the form of a frequency distribution, a *mode* is a value to which corresponds a greater frequency than to values just preceding or values immediately following it. A frequency distribution may then have more than one mode, although we shall confine our attention almost wholly to frequency distributions which have only one mode.

The great service of the mode is to characterize a type. Thus, when we say that a certain man is an average citizen we mean that he represents a type which is met oftener than any other; we certainly do not refer to an arithmetic average or a median. Thus, if a few citizens of a community are millionaires, and all the rest, to quite a number, are in poverty, we should say that the average citizen is in poverty, meaning by average citizen the "modal" citizen; an arithmetic average of the wealth of the community might give the erroneous impression that the people of the community are in good financial condition. The mode as an average suffers from the fact that it can not usually be determined with any great degree of accuracy—at least, from the average frequency distribution.

40. Rectangular Histograms: Histograms of Three Dimensions.—If, in plotting the points corresponding to the observations and frequencies of a distribution, the ordinates were replaced by vertical rectangles of proportionate length and of width proportional to the class interval, where the midpoint of the base of a rectangle is taken at the corresponding class mark as an abscissa, the aggregate of rectangles—called *frequency rectangles*—is called a *rectangular histogram.* The histogram corresponding to frequency distribution No. 2, Art. 38, would appear somewhat as follows:

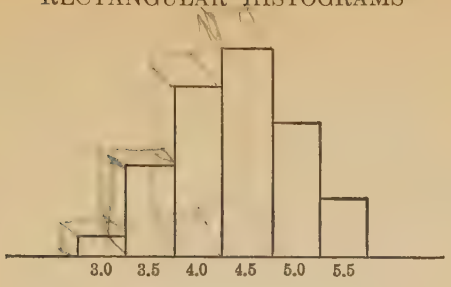


FIG. 1.

If the ordinates of the graph of an analytic expression corresponding to successive integral values of the abscissas were drawn, and then the corresponding histogram were drawn, the area of the histogram would be given by ordinary summation or finite integration. Thus, the geometrical interpretation of finding the sum of a finite number of terms of a series is the determination of the area of the corresponding histogram. The area under the corresponding curve would evidently be given by ordinary integration. The area of the histogram would, however, be only approximately equal to the area under the corresponding curve. At least a slight discrepancy between the two areas is then to be expected even when the analytic expression is known. As this analytic expression is rarely known in practice, another source of discrepancy would lie in the choice of the analytic expression selected to fit the observed frequencies; that is, the analytic expression selected could rarely be expected to fit a given frequency distribution exactly. The relation between the various characteristics of a histogram and the corresponding frequency curve determined analytically forms the essential basis for the explanation of the discrepancies between results obtained directly from statistical data and results obtained by the formal processes of mathematical analysis. The effort should be made simply to control the values of such discrepancies sufficiently to obtain the desired accuracy in the final results.

Frequency distributions considered heretofore are said, for

obvious reasons, to be of two dimensions. Frequency distributions of three dimensions are also common and are usually given in the following form, which gives the results of an investigation into the relation between the grades by psychological test and mathematical grades of 307 students:

Mathematical Grades	Psychological Test											
	95	105	115	125	135	145	155	165	175	185	195	Total
95	...	....	1	....	1	2	5	7	8	4	1	29
85	...	....	2	1	1	7	10	10	8	3	1	43
75	...	....	8	16	23	35	28	22	15	5	....	152
65	...	1	6	7	3	7	8	7	1	3	....	43
55	1	3	3	6	8	6	6	6	1	....	....	40
Total....	1	4	20	30	36	57	57	52	33	15	2	307

It will be noticed that the totals at the right and at the bottom constitute frequency distributions of mathematical grades and grades by psychological test respectively.

If parallelopipeds were thought of as erected upon the xy plane of lengths proportional to the frequencies and corresponding to the class marks given in such a table, the aggregate of parallelopipeds would form a histogram of three dimensions; and if a surface were thought of as fitting the frequencies or the histogram somewhat as a frequency curve fits ordinates representing frequencies or a rectangular histogram, the surface is called a *frequency surface.*

41. Deviations: A Method of Computing the Arithmetic Average.—We propose now to illustrate how the formation of frequency distributions may be employed to simplify the computation of the arithmetic average. It is desirable first, however, to introduce also the idea of *deviations.*

Suppose that we select a *trial*, or *provisional*, mean of a given distribution by inspection and denote it by M'. Then

the average d, of the deviations of all the observations of the distribution, say $x_1, x_2, \ldots x_n$ from this trial mean, is obviously

$$d = \frac{(x_1 - M') + (x_2 - M') + \ldots + (x_n - M')}{n} = M - M',$$

or

$$M = M' + d, \quad \ldots \ldots \ldots \ldots \ldots \quad (36)$$

where M denotes the true mean. Formula (36) then shows that if we work with the deviations from a trial mean instead of with the original observations the average so obtained, or d, constitutes a correction which when added algebraically to the trial mean will give us the true mean. A few examples will verify the fact that such a plan simplifies the computation because large factors will thereby be replaced by smaller factors. It is easily verified that the average of the deviations from the true mean must be zero.

Computation of the mean will be further simplified if we replace the unit of measurement at first by unity and replace it later, in which case formula (36) may evidently be written

$$M = M' + kd', \quad \ldots \ldots \quad (37)$$

where k is the unit of measurement and d' is the average of the deviations from the trial mean when this unit of measurement is taken to be unity. The application of formula (37) is shown below.

Original Observations	Frequencies y	Deviations x	xy
3.0	1	−3	− 3
3.5	5	−2	−10
4.0	9	−1	− 9
4.5	11	0	
5.0	7	1	7
5.5	3	2	6
	36		− 9

The trial mean is $M' = 4.5$;
the unit of measurement is $k = 0.5$;

and the unit correction $d' = -\frac{9}{36} = 0.25$.

Hence, the true mean is $M = 4.5 + 0.5(-0.25) = 4.375$ or 4.38.

Many observations appear originally in the form of deviations, and such deviations constitute errors of observation so frequently that the term "errors" is used very widely as synonymous with the term "deviations." The term "error" then implies far more in such a connection than the term usually implies.

EXERCISES

1 and **2.** Form frequency distributions of the data of Exercises 1 and 2 of the preceding list and compute the arithmetic averages.

(*a*) Form a frequency distribution of the data of Exercise 2 with class marks 3, 8, 13, 18, etc., and compute the arithmetic average.

3. A set of examination books in geometry were subjected to two independent readings by the College Examination Board to give the following differences in credit (on a basis of 100) and corresponding frequencies:

Differences	0	1	2	3	4	5	6	7	8	9
Frequencies	5	5	7	9	5	14	2	8	6	5
Differences	10	11	12	13	14	15	18	20	22	31
Frequencies	4	1	4	2	1	3	2	2	1	1

(*a*) Compute the average difference. *Ans.* 7.01.

(*b*) Construct a frequency distribution with class marks 1, 4, 7, 10, etc., and compute the arithmetic average. *Ans.* 7.00.

4. Another example of the same kind as the preceding yielded:

Differences	0	1	2	3	4	5	6	7	8	9
Frequencies	6	5	22	12	3	15	2	8	0	3
Differences		10	11	12	13	14	15	16	17	18
Frequencies		3	1	0	2	0	3	0	1	1

Follow the instructions given in the preceding exercise.

Ans. 4.76.
4.62.

5. The average credit received by all students in one year at a certain college was as follows:

Credit	Frequency	Credit	Frequency
4.05	12	1.65	290
3.65	34	1.25	174
3.25	96	0.85	81
2.85	128	0.45	33
2.45	244	0.05	9
2.05	250		

Compute the average credit. What single credit (i.e., the mode) is most likely?

6 and **7.** The following frequency distributions give, first, the lengths of 800 ears of corn in inches, and second, the heights of the freshmen at a certain university.

Lengths of Corn	Frequencies	Heights (inches)	Frequencies
4.0	1	61	2
4.5	1	62	10
5.0	8	63	11
5.5	33	64	38
6.0	70	65	57
6.5	110	66	93
7.0	176	67	06
7.5	172	68	26
8.0	124	69	09
8.5	61	70	87
9.0	32	71	75
9.5	10	72	23
10.0	2	73	9
		74	4

(*a*) Compute the average length (or height).

(*b*) Form a frequency distribution with class marks 4.25, 5.25, 6.25, etc. (61.5, 63.5, 65.5, etc.), and compute the average length (or height).

(*c*) What length (or height) was most common?

8. The weights of the freshmen of a certain university were found to be as follows:

Pounds	Number	Pounds	Number
104.5	21	154.5	87
114.5	68	164.5	35
124.5	169	174.5	14
134.5	203	184.5	11
144.5	142		

Compute the average weight.

9. The following data are the partial results of an investigation into the cost of living in the District of Columbia for 1917, among representative families drawing salaries of $1800 or less (incomes given in the table which exceed $1800 refer to families which received remuneration in addition to salaries).

Income	Average Size of Family	Number of Families
$300	2.2	10
500	3.1	54
700	3.5	156
900	3.8	247
1100	3.7	242
1300	4.0	280
1500	3.9	221
1700	4.0	122
1900	4.1	87
2100	4.2	35

Assign a trial mean, ignoring the unit of measurement until the last, and

(*a*) Compute the average income.

(*b*) Compute the average size of a family.

(*c*) What size of family and what income were most likely?

10. The following data give the ages at which infection of leprosy was "supposed" to have taken place among cases investigated in Hawaii.

Ages	Frequency	Ages	Frequency
1– 5	8	46–50	48
6–10	56	51–55	41
11–15	163	56–60	26
16–20	204	61–65	18
21–25	143	66–70	18
26–30	114	71–75	3
31–35	79	76–80	3
36–40	89	81–85	1
41–45	44		

Determine the average age of infection.

At what single age was infection most likely?

42. Normal Distributions.—It is needless to say that egregious blunders can upset any kind of an investigation and that we must take their absence for granted. Errors of smaller size are, however, inevitable and are of the greatest importance in the analysis of statistics. A vast majority of these errors are compensating in character and tend, in the long run, not only to hover or concentrate about the mean but also to occur with the same frequency corresponding to each size of deviation on one side of the mean as on the other. Such distributions are called *normal* distributions and will be given much consideration in later sections. It will be desirable for the present to include as normal distributions many distributions which possess discrepancies which seem to be due to lack in the number of observations and which would promise to disappear if the number of observations were increased indefinitely. The graph of a truly normal distribution would then be symmetrical with respect to an axis of reference erected at the mean.

There will be found many frequency distributions which differ markedly from a normal distribution. However, the great majority of all the distributions will prove to be normal, according to the loose definition given above, when we have cogent reason for believing that deviations of any size on one side of the mean are as likely as deviations of the same size on the other side.

We propose to assume that distributions of errors of observation would prove to be normal whenever we have cogent reason for the assumption, and we usually have cogent reason for the assumption when the observations refer to the size of a single object, such as the deviations from the most probable value of a large number of readings of a barometer made under the same conditions. It should, of course, be kept in mind in that connection that the actual distributions of errors of observation will usually be entirely lacking and that the assumption stated above is the best at our disposal.

Frequency distributions of measurements of various objects, even though these objects belong to the same particular class, may or may not be normal. Thus, the distribution of the heights of a large group of individuals all of the same age might be fairly normal, but the distribution of the individuals of a community with respect to wealth would be pretty sure to differ widely from a normal distribution. In neither of the last two illustrations would we have cogent reason for assuming the distributions to be normal.

Even distributions of measurements made upon a single object differ occasionally from normal distributions for some particular reason which may not be suspected at first thought. Thus, if we set up a frequency distribution of the results of estimating the center of a book with the point of a knife blade, our knowledge of the inequality of the strength of the two eyes would probably affect any cogent reason for expecting a normal distribution.

43. The Standard Deviation or Dispersion.—One variation from the method of computing an average, such as the arithmetic average or mean of a set of observations, would be to square each deviation from the mean, determine the average of the squares, and then extract the square root of this average. The final result is called by the expressive term *root mean square*, or more generally, the *standard deviation* or *dispersion* of the observations. As an example, the squares of the devia-

tions of the observations given at the beginning of this chapter from their mean 4.38 would be

2.62	0.38	0.05	0.00	0.12	0.38
1.12	0.34	0.05	0.00	0.12	0.38
0.49	0.21	0.05	0.00	0.14	0.49
0.44	0.21	0.01	0.05	0.25	0.81
0.44	0.09	0.00	0.07	0.34	1.04
0.44	0.07	0.00	0.09	0.38	1.80

It is easily verified that the average of these squares is 0.374 and that the value of the standard deviation is then 0.61.

Special attention is called to the fact that the largest deviations are given special emphasis in computing the value of the standard deviation, since they are squared before an average is taken, and numbers greater than unity are increased and numbers less than unity are diminished by the process of squaring. This is usually very desirable because the presence of large deviations is usually very important. The value of the standard deviation constitutes a good measure of the consistency of a set of observations or the extent to which the observations differ from the mean. Hence, we are usually very much interested in any inconsistencies which are apt to affect our confidence in the value of the mean. The method of computing the standard deviation obviates, then, any possibility of overlooking any such inconsistencies. As an example, suppose that two groups of individuals were to make readings of a finely graduated barometer. If the standard deviations of the two sets of observations should differ significantly, we would naturally place more reliance in the mean of the set whose standard deviation had the less value, because that set of observations would be regarded as more consistent.

The standard deviation or dispersion will prove to be of great fundamental importance in almost all of the work of the succeeding chapters. It is well, then, that we look for a much easier method of computing it. Such a method, like that of computing the mean, is based upon the use of a frequency dis-

tribution of deviations from a trial mean. Here, again, we shall find that the compression of a set of observations into a frequency distribution, with a corresponding sacrifice in the accuracy of the expression of the individual observations, affects much less the accuracy of an average such as the standard deviation.

If we denote the square of the dispersion of a set of observations $x_1, x_2, \ldots x$ about a trial mean M' by s^2 then

$$s^2 = \frac{\Sigma(x_i - M')^2}{n} = \frac{\Sigma(x_i - M + M - M')^2}{n} = \frac{\Sigma(x_i - M + d)^2}{n}$$

$$= \frac{\Sigma(x_i - M)^2}{n} + 2d\frac{\Sigma(x_i - M)}{n} + d^2,$$

where d denotes the correction to be applied to the trial mean M' to give the true mean M.

But $\frac{\Sigma(x_i - M)}{n}$ and $\frac{\Sigma(x_i - M^2)}{n}$ are, respectively, the average of the deviations from the mean and the square of the dispersion which we desire, of which the value of the first is zero. If, then, we denote the dispersion about the mean by σ we have

$$s^2 = \sigma^2 + d^2,$$

or

$$\sigma^2 = s^2 - d^2. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (38)$$

It is left for the student to verify from the following example that if the unit of measurement is k and the standard deviation with this unit of measurement replaced by unity is σ' then

$$\sigma = k\sigma'. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (39)$$

The application of formulas (38) and (39) are shown below in connection with the very much compressed frequency distribution given previously:

Observations (class marks)	Frequencies y	Deviations x	xy	x^2y
3.0	1	−3	− 3	9
3.5	5	−2	−10	20
4.0	9	−1	− 9	9
4.5	11	0		
5.0	7	1	7	7
5.5	3	2	6	12
	36		− 9	57

Hence

$$d=\frac{-9}{36}=-0.25,$$

$$s^2=\frac{57}{36}=1.58,$$

$$\sigma'^2=1.58-0.06=1.52,$$

and

$$\sigma=k\sigma'=0.5\sqrt{1.52}=0.76.$$

It will be recalled that the value of the standard deviation obtained directly from the original observations was 0.61.

44. Average Deviation.—We have already referred to the fact that it is characteristic of the standard deviation to give special emphasis to the larger deviations. Cases arise occasionally where this emphasis is unnecessary or even undesirable. Under such circumstances the *average deviation* usually proves useful and may be defined as the arithmetic average of the *absolute values* of the deviations from the mean. It is easily verified that the average deviation of the observations given at the beginning of this chapter is 0.49, and that the average deviation of the deviations of the frequency distribution considered in the preceding section is also 0.49. It should be noted that this value is less than that found for the standard deviation, for the reason stated above.

EXERCISES

1 and **2.** Compute the values of the standard deviation of the distributions given in Exercises 3 and 4 of the preceding list. The

second distribution of that list was obtained after a somewhat more thorough examination than in the case of the first distribution. Do the values of the standard deviation reflect this fact? Show that the value of the ratio of the standard deviation and the mean is greater for the second distribution than for the first. Compute the average deviations. *Ans.* 4.25 and 3.05.

3. Scores in whist, by evenings, were made by two players as follows:

A: −51, 19, 8, −23, 15, 33, 11, −1, 3, 35, −84, 34, −20, 42, 41, −33
B: 5, −37, 5, −20 66, −84, 16, −3, −15, −83, 38, 20, −5, 31, 17, 47

These scores are expressed in sevenths to avoid fractions. Take this fact into consideration and compute the actual average score and dispersion of each player. Which player has the better average and which appears to be more consistent?

4. The scores of a famous cricketer for two years were as follows:

1905: 27, 76, 14, 13, 47, 45, 39, 7, 15, 34, 38, 106, 107, 75, 8, 3, 4, 4, 4, 47, 13, 209, 66, 14, 0, 78, 10, 23, 0, 0, 44, 16
1906: 0, 10, 204, 0, 0, 60, 0, 85, 0, 49, 86, 3, 50, 5, 10, 7, 70, 43, 9, 59, 97, 0, 166, 152, 13, 22, 34, 82, 14, 14, 35, 56, 48, 23, 31

Compute the average and the dispersion for each year.

Ans. 37.1, 43.1; 43.9, 49.3.

Note that the value of the ratio of the dispersion and the mean was greater for 1905 than for 1906.

5. Follow the instructions given in the preceding exercise but for the scores:

1906: 9, 37, 24, 24, 58, 10, 22, 255, 13, 73, 1, 5, 47, 3, 110, 80, 0, 16, 8, 21, 20, 3, 31, 70, 55, 26, 60, 68, 32, 143, 80, 5, 19, 1, 115, 82, 70, 52, 6, 85, 6, 33, 31, 6, 59, 2, 66, 71, 63, 44, 8, 40, 122, 29, 38, 76, 32, 17
1907: 39, 82, 25, 5, 219, 135, 0, 46, 0, 68, 10, 31, 194, 244, 143, 125, 54, 69, 6, 15, 100, 144, 34, 144, 208, 54, 34, 76, 10, 16, 10, 109, 51, 18, 14, 76, 168, 17, 3, 94, 45, 51, 8, 9, 24, 60, 110, 141, 22, 28, 62, 51, 17. *Ans.* 44.5, 43.9; 66.4, 62.0.

45. The Coefficient of Variation.—In comparing the way two things vary it should be evident that relative size influences not only the mean but also the deviations from it. Stating the matter a little differently, if the mean is greater in one investigation than in another it is only natural to expect the deviations from the mean to be greater. A better measure of the variability of a character for purposes of comparison is usually given, therefore, by dividing the value of the standard deviation by the value of the mean. The *coefficient of variation* v is defined by the relation

$$v = \frac{100\sigma}{M},$$

where M denotes the mean.

The formula for the coefficient of variation does not fit very easily into the many formulas which will be considered in the following chapters, and it is rarely necessary in extended investigations, such as will be treated later, to subject the variability of characters to an analysis beyond that afforded by the value of the standard deviation alone. However, if the main object of an investigation is to measure the variability of a character, and accuracy is essential, then the value of the coefficient of variation should also be computed.

As an example, let us compare the records of the famous English cricketer, Hayward, for the years 1905 and 1907, and determine in which year he was more consistent. The scores for these years are as follows:

1905: 43, 27, 13, 9, 10, 34, 59, 53, 13, 116, 11, 4, 3, 31, 22, 35, 3, 24, 21, 128, 168, 52, 122, 17, 148, 91, 88, 14, 203, 13, 64, 177, 88, 76, 106, 112, 2, 25, 48, 64, 26, 24, 216, 81, 58, 14, 33, 32, 35, 53, 10, 9, 197, 12, 28, 28, 0, 44, 2.

1907: 39, 82, 25, 5, 219, 135, 0, 46, 0, 68, 10, 31, 194, 244, 143, 125, 54, 69, 6, 15, 100, 144, 34, 144, 208, 54, 34, 76, 10, 16, 10, 109, 51, 18, 14, 76, 168, 17, 3, 94, 45, 51, 8, 9, 24, 60, 110, 141, 22, 28, 62, 51, 17.

The following results are obtained:

	Average Score	Standard Deviation	Coefficient of Variation
1905	54.9	55.0	100
1907	66.4	62.0	93

It follows, then, that although the value of the standard deviation of the scores was greater for 1907 than for 1905, we should be unjustified in this case to conclude that Hayward was more consistent in 1905, for his average was so much greater in 1907 that the value of the coefficient of variation was less for that year than for 1905.

EXERCISES

1. The scores in whist made by seven players, by evenings, are as follows:

A	B	C	D	E	F	G
13	12	18	−37	− 9	− 9	12
−35	65	−11	8	−17	−27	65
10	16	−15	17	18	6	16
− 8	− 8	23	−11	11	−11	− 6
48	48	37	−19	− 1	−19	17
7	7	3	8	3	−28	−20
41	8	17	7	−16	7	17
5	−18	− 5	−14	5	21	5
4	44	−32	−29	37	−34	−32
34	8	8	1	−16	11	34
22	−15	−15	23	−14	23	22
13	15	15	− 4	0	−33	13
29	−28	16	−23	1	−12	11
52	21	52	− 9	−11	− 2	35
−14	11	−25	41	11	−13	−14
− 5	−36	14	− 5	−17	− 8	15

These scores are expressed in fifths to avoid fractions. Take this fact into consideration and compute the actual average score and dispersion of any two players. Note not only which player has the better average, but also whether the ranking of the players in regard to consistency is affected by the size of the average score.

2. The exports of the United States and of Great Britain, in millions of dollars, for the first eleven months of 1921 were as follows:

	United States	Great Britain
January	279	302
February	251	298
March	330	327
April	318	285
May	308	298
June	335	271
July	301	305
August	302	301
September	313	305
October	371	305
November	383	339

Compute and compare the values of the dispersions and coefficients of variation.

3. Two hundred estimates of the center of a book (a different book was used in each case and the ends of the book were reversed after each estimate) were made by each of four individuals to give the following distributions:

A's		B's		C's		D's	
Page	Frequencies	Page	Frequencies	Page	Frequencies	Page	Frequencies
480	4	545	1	540	2	615	1
485	21	550	1	550	10	625	1
490	8	555	4	560	10	635	10
495	15	560	3	570	24	645	25
500	19	565	7	580	23	655	39
505	33	570	18	590	28	665	46
510	23	575	21	600	34	675	32
515	17	580	17	610	25	685	27
520	12	585	28	620	21	695	10
525	19	590	22	630	13	705	7
530	13	595	31	640	9	715	2
535	11	600	14	650	1		
540	1	605	13				
545	2	610	11				
550	1	615	3				
555	1	620	4				
		625	1				
		630	1				

Compute the coefficients of variation and rank the estimators in the order of their consistency. v

Ans. D, 0.27.
C, 0.40.
B, 0.51.
A, 0.62.

4. The scores of two cricketers for three seasons were as follows:

A: 0, 0, 4, 20, 61, 26, 8, 87, 10, 29, 10, 52, 5, 27, 40, 12, 2, 11, 34, 206, 48, 43, 0, 1, 1, 3, 7, 27, 77, 17, 43, 42, 24, 20, 23, 10, 68, 170, 58, 47, 77; 28, 1, 60, 38, 14, 48, 0, 0, 1, 25, 1, 17, 20, 11, 14, 89, 234, 5, 23, 0, 1, 52, 4, 2, 1, 4, 19, 15, 54, 21, 65, 57, 60, 8, 10, 15, 11, 1, 45, 0, 74, 8, 25, 0, 43, 0, 1; 63, 55, 5, 17, 16, 42, 8, 8, 87, 27, 22, 34, 4, 48, 5, 80, 66, 40, 0, 15, 6, 12, 111, 0, 4, 19, 53, 4, 6, 0, 48,15, 75, 26, 28, 1, 10, 34, 2, 14, 9, 54.

B: 27, 76, 14, 13, 47, 45, 39, 7, 15, 34, 38, 106, 107, 75, 8, 3, 4, 4, 4, 47, 13, 209, 66, 14, 0, 78, 10, 23, 0, 0, 44, 16; 0, 10, 204, 0, 0, 60, 0, 85, 0, 49, 86, 3, 50, 5, 10, 7, 70, 43, 9, 59, 97, 0, 166, 152, 13, 22, 34, 82, 14, 14, 35, 56, 48, 23, 31; 61, 42, 137, 10, 41, 72, 0, 22, 53, 66, 4, 73, 28, 17, 16, 122, 48, 5, 87, 26, 9, 66, 7, 9, 17, 44, 60, 12, 66, 2, 77.

Compute the values of the coefficients of variation and show that although B's standard deviation is larger his coefficient of variation is smaller. A's $v=124$.
B's $v=105$.

46. Some Applications of Averages with Time Series. Index Numbers.—The standard deviation has been included among the various averages in this chapter but it is more properly a measure of dispersion. If we omit it from consideration then for the time the arithmetic average or mean is easily the most important form of an average not only because it is most widely known and used but also because it is definitely connected with a more comprehensive scheme (including dispersion, etc.) for handling statistical data to give more reliable results *when these data are satisfactory in number and quality.* Statistical data, however, are so frequently insufficient and so roughly approximate that a certain amount of guesswork in

choosing the most representative average is often desirable and in many such cases there are clear evidences that some other average, particularly the median, should be selected. Several illustrations of the latter fact will be given as we proceed.

One of the oldest forms of statistical data consists of data—prices, production, etc.—arranged by periods of time (that is, by months, years, etc.) and is usually referred to as a *time series*. A very valuable kind of investigation is the analysis of a time series to measure changes in commodity prices and thereby enable us to foresee to some extent important periods of prosperity or of depression. In presenting the most widely accepted form of such an analysis it will evidently be necessary to omit here all considerations of the possibly very important effect of the change in the intrinsic value of the monetary unit, say the gold dollar.

Let us consider the time series of the prices in this country of an important commodity such as cotton. The first few terms of such a series beginning with the year 1895 is as follows:

Year	Cents per Pound
1895	7.30
1896	7.92
1897	7.15
1898	5.97
1899	6.58
etc.	

The trend of such a series can be indicated more forcibly by choosing a *base year* and dividing all the prices by the price for that year—called the *base price*—to give ratios called *price relatives*. Price relatives are usually *stated* in percentages but should be used in the simple form of ratios for further computation. If instead of using a fixed base price we use in each case the price in the preceding year as base we obtain ratios known as *link relatives*. Link relatives are also usually *stated* in percentages. The following items illustrate sufficiently the two forms of relatives:

Cotton

Year	Price	Relative to 1895	Link Relative
1895	7.30	100.0	
1896	7.92	108.5	108.5
1897	7.15	97.9	90.3
etc.			

Another form of relative known as the *chain relative* is frequently used and is formed from the column of link relatives by assigning 100 as the chain relative for the first year and defining the chain relative of any other year as the product of the link relative for that year and the chain relative for the preceding year. It should be clear, however, that such a process applied to the table given above would simply reproduce the column of fixed base relatives. We shall see later that that statement would be true only so long as we restrict our investigation to the prices of a single commodity.

We come finally to the most valuable and at the same time the most natural means of studying and analyzing such a time series—its graph. In a non-mathematical course much time and stress would now be devoted to this important part of the analysis but we shall be obliged to pass on to less familiar steps in our treatment. Before we pass on, however, it should be noted that such graphs are valuable mainly in showing the *amount* of change in price at various times. It is frequently more important to compare the *rates* of change in price at various times and since it is naturally easier for the eye to measure segments rather than slopes, a more appropriate graph for that purpose would be the graph of the values of the logarithms of the prices or better a graph of the prices themselves but plotted on logarithmically ruled paper.

Perhaps we have unintentionly emphasized the year as the unit period of time. Let us think then of the processes explained above as applied more particularly to prices or data by months and let us compare the graphs of prices of several commodities. We shall find the graphs to vary so much that it is unsafe to

draw any conclusions other than those which are definitely connected with the particular commodities under consideration at the time. As we are much more concerned as a general rule with the trend of all prices a much more reasonable procedure would be to work with averages of prices or averages of price relatives of a large number of commodities on the assumption that a summary of the individual movements of the prices of commodities taken as a group will reflect analogous movements in the general price level. Such averages are called *index numbers*. It will be seen very shortly, if not at once, that there might well be no justification for this assumption unless care is exercised to combine or weight properly the prices which are used to compute the values of the index numbers.

As a matter of fact, the process of weighting is a double one. That is, it will be necessary first to eliminate false weights which would naturally be present originally due to differences in prices or, what amounts to the same thing, to different units of measurement of quantity, regardless of the relative importance of the commodities, and second, to apply new weights which will reflect the relative importance of the commodities. Thus, an index number based upon the actual prices of coal per ton and cotton per pound would clearly be absurd.

The original improper weighting can be eliminated simply by using the price relatives of the various commodities and their averages. Let us take the series of prices of each of eight of the most important commodities which affect the cost of living.

	Prices							
	Wheat (bu.)	Cattle (cwt.)	Hogs (cwt.)	Cotton (bale)	Hard Coal (ton)	Soft Coal (ton)	Petrol (bbl.)	Pig Iron (ton)
1913	0.986	8.507	8.454	0.128	5.061	2.413	2.450	14.71
1914	1.005	9.039	8.382	0.121	5.061	2.400	1.917	12.87
1915	1.307	8.702	7.187	0.102	5.045	2.450	1.529	13.74
1916	1.351	9.573	9.400	0.145	5.454	3.067	2.483	19.76

etc.

The corresponding price relatives with 1913 as the base year are as follows:

1914	101.9	106.2	99.1	94.6	100.0	99.5	78.2	87.5
1915	132.5	102.3	85.0	79.7	99.7	101.6	62.4	93.4
1916	136.9	112.5	111.2	113.1	107.8	127.1	101.4	134.4

etc.

It is well to pause at this point and decide upon the form of an average which should be used. The values of the various kinds of averages of the relatives for the specified years are as follows:

	1913	1914	1915	1916
Arithmetic	100.0	95.9	94.6	118.1
Median	100.0	99.3	96.6	112.8
Harmonic	100.0	95.1	90.7	116.8
Geometric	100.0	95.5	92.6	117.4

The use of the mode with so few commodities is, of course, out of the question. The choice of an average should always depend upon a preliminary inspection and analysis and no fixed rules can be laid down except that that average should be chosen which would seem to be most representative. With so few cases it is probably the best plan to arrange the variates so that some idea of the dispersion can be gained by inspection; if there is no great disparity the arithmetic average will probably be the best, otherwise the median must be given due consideration. Strangely enough, the harmonic mean becomes very important when we undertake to apply proper weights, for reasons to be explained.

The necessity of both unweighting and reweighting prices for forming satisfactory index numbers should be evident. Thus, it would be absurd to expect to obtain a representative index number from the actual prices of wheat and diamonds but it should be equally evident that the relatives for these two commodities should be weighted to reflect the greater importance of wheat in investigating changes in the cost of living. There are various ways of measuring the relative importance of commodities but most of them reduce somewhat to that of

determining the relative amounts of production whether determined through values of exchange, estimates of consumption, direct estimates of production or any other source, all of which frequently fail to give more than rough approximations of the quantity of a given commodity produced.

If weights are selected proportional to the quantities of production of the commodities and if these weights vary from period to period it is known that the index numbers so obtained are affected by an upward *bias* (and a downward bias if a fixed set of weights is used). Moreover, there is usually a bias according to the form of average [1] used (except the simple geometric average). The simple arithmetic average tends to give excessive results and the simple harmonic mean tends just the opposite direction. One of the most effective ways of eliminating bias has been found to *cross* index numbers having opposite bias. As a proper presentation of an effective method of crossing indexes would call for a more extensive treatment and discussion than can be properly accorded here we shall conclude this discussion by simply quoting what is called Fisher's *Ideal Formula*,[1] which is a geometric average of an arithmetic index of relatives with fixed (base) year values as weights and a harmonic index of relatives with variable year values as weights.

$$I = \sqrt{\frac{\Sigma p_0 q_0 \left(\frac{p_i}{p_0}\right)}{\Sigma p_0 q_0} \times \frac{\Sigma p_i q_i}{\Sigma p_i q_i \left(\frac{p_0}{p_i}\right)}},$$

or simply

$$I = \sqrt{\frac{\Sigma p_i q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_i q_i}{\Sigma p_0 q_i}},$$

where p_i and q_i denote the price and the quantity respectively in any period and p_0 and q_0 the base price and the base quantity respectively. It should be noticed that $p_i q_0$ may be obtained by multiplying the relative p_i/p_0 by the base value $p_0 q_0$. It

[1] See Irving Fisher's *The Making of Index Numbers* for an extensive discussion.

should be added that certain *tests* should be applied in practice to determine the reliability of the results given by a formula for index numbers such as Fisher's Ideal formula, but the student must be referred elsewhere [2] for an explanation of these tests as well as a more detailed treatment of index numbers and time series—with useful exercises. The consideration of time series with further applications of averages will be continued in Chapter VII in connection with forecasting of business conditions.

EXERCISES

1. Given the following yearly average prices of corn per bushel:

1895	39.6	1900	38.1	1905	50.1
1896	25.8	1901	49.7	1906	46.3
1897	25.5	1902	59.7	1907	52.8
1898	31.4	1903	46.1	1908	68.4
1899	33.3	1904	50.5	1909	66.8

make tables of values of fixed base (1895 as base) relatives and of link relatives.

2. Given the following estimates of production, employ the table of prices given in the text to find the values of the index numbers for the specified years by Fisher's Ideal formula.

	Wheat	Cattle	Hogs	Cotton	Hard Coal	Soft Coal	Petrol	Pig Iron
1913	763.4	94.0	107	14.16	91.52	478.4	248.4	30.97
1914	891.0	91.0	102	16.14	90.82	422.7	265.8	23.33
1915	1025.8	96.5	120	11.19	89.00	442.6	281.1	29.92
1916	636.3	112.0	135	11.45	87.58	502.5	300.8	39.44

[2] Crum and Patton, *Introduction to Methods of Economic Statistics*. This publication includes also a good account of the various current systems of index numbers employed by well known authorities. Current developments in the whole field of time series will be found in the successive issues of the *Harvard Economic Service*.

CHAPTER VII

MOMENTS

47. Moments of Frequency Distributions and Curves.— We have noticed that when a large number of careful observations are made, and when these observations are classified to form a frequency distribution, the frequencies are usually distributed in such a smooth manner that frequency curves are naturally suggested. These frequency curves have been studied and investigated at great length in various ways, but in particular by what is called the method of *moments*. Although moments are employed in various branches of applied science, it will be found desirable to develop the method along special lines in order that it may be applicable to the discrete form of statistical data.

We shall define the n-th moment of a point (x, y) with respect to the y-axis as the product of the ordinate y and the n-th power of the abscissa x. We may then extend this definition to certain strips of area by the relation

$$n\text{-th moment} = \lim_{\Delta A \to 0} \Sigma x^n \Delta A, \quad . \quad . \quad (40)$$

where ΔA is an element or strip of the given area parallel to the y-axis whose distance from the y-axis is x, and where the summation is to extend over the entire area concerned. It follows, then, that if the area is bounded by a curve whose equation is $y = f(x)$, the ordinates at $x = a$ and $x = b$, and the x-axis, relation (40) may be replaced by the definite integral

$$n\text{-th moment} = \int_a^b x^n f(x) dx. \quad . \quad . \quad . \quad . \quad (40')$$

If the area under consideration is that of a rectangular histogram, relation (40) may be written in the more suitable form

$$n\text{-th moment} = \sum_{x=a}^{b} x^n y \quad . \quad . \quad . \quad . \quad . \quad . \quad (40'')$$

which naturally suggests the use of finite integration. Finite integration is singularly appropriate when no simpler method is available; but so many valuable and instructive applications can be made where ordinary multiplication and addition will suffice that we shall confine our attention for the most part to those applications. Certain applications which call for the use of the integral calculus, and which will lead to some of the most important conceptions treated in this course, will be suggested in exercises and in the theory considered in later sections. The following small distribution gives the frequencies of litters of the corresponding number of mice as found in certain breeding experiments. The typical method of computing the first and second moments is shown to the right. The higher moments are computed in a similar manner.

Number in Litter x	Frequency y	xy	x^2y	$(x+1)^2y$
1	7	7	7	28
2	11	22	44	99
3	16	48	144	256
4	17	68	272	425
5	26	130	650	936
6	31	186	1116	1519
7	11	77	539	704
8	1	8	64	81
9	1	9	81	100
	121	555	2917	4148

Thus, the zero-th moment or total frequency = 121,
the first moment = 555,
and the second moment = 2917.

To the right is shown a method of checking the computation of the first and second moments, called the *Charlier* method of check. Since $(x+1)^2y = x^2y + 2xy + y$, it is evident that the value of $\Sigma(x+1)^2y$ must be the same as the sum of the second moment, *twice* the first moment and the zero-th moment. In this case

$$4148 = 2917 + 2(555) + 121.$$

EXERCISES

Compute the values of the first and the second moments of the following distributions:

1. The frequencies of the various numbers of certain glands found in the right forelegs of 2000 female swine were found to be as follows:

Number	Frequency	Number	Frequency
0	15	6	134
1	209	7	72
2	365	8	22
3	482	9	8
4	414	10	2
5	277		
			2000

What is the value of the mean?

2 and **3.** The frequencies of the various numbers of children per wife were compiled from certain genealogical records of American families for several periods, of which the following distribution (2) refers to the period preceding 1700 and the distribution (3) to the period 1870–1879:

Number of Children	Frequency (2)	Frequency (3)
0	5	88
1	5	202
2	11	265
3	9	210
4	11	147
5	29	77
6	34	51
7	26	22
8	47	11
9	32	9
10	29	4
11	19	
12	11	
13	4	
14	3	
15	0	
16	1	

What is the average number of children per wife?

4. The budgets of 421 Smith College girls for 1914–1915 were found to be:

$450	66
650	169
850	109
1050	43
1250	20
1450	8
1650	3
1850	3
	421

What is the average budget?

48. Fitting Curves by Moments.—We have already mentioned the frequent desirability of fitting curves to given frequencies, that is, of determining the analytic expression that gives computed frequencies which agree very closely with the given frequencies. We have mentioned also the peculiar ease with which rational integral functions may be used for this purpose. Moments are very useful in fitting rational integral functions, or polynomials of the form $y=a+bx+cx^2+$ etc., to given frequencies.

In fitting such curves it is well to decide beforehand, from the appearance of the frequency curve obtained by plotting the given frequencies, what degree of the polynomial seems most appropriate with respect to both accuracy of fit and simplicity of the polynomial desired. It should be kept in mind that it is always possible to find the equation of a polynomial of the n-th degree which is satisfied exactly by the coordinates of any $n+1$ points. Such a curve, however, may fit the situation less well than a simpler curve, because many, if not most, of the vagaries of the positions of the points may be due to errors in observation. In the problem before us, it is not to be expected that the curve finally obtained will fit exactly all the points corresponding to the various frequencies, but merely

that the curve will fit them as well or better than any other curve of the same kind and degree.

The fundamental line of procedure is to equate moments of the theoretical curve (represented by the polynomial selected) to the corresponding moments of the given frequencies and thereby set up equations whose solution will give the values of the coefficients of the polynomial. The method can be explained best by a simple example. Let us fit the straight line $y=a+bx$ (that is, determine the values of a and b) to the points (1, 2), (3, 9) and (5, 14). It is obvious on plotting that no straight line will pass exactly through the three points.

The zero-th moment of the polynomial is the sum of the ordinates or values of y for $x=1$, 3 and 5 or

$$(a+b)+(a+3b)+(a+5b)=3a+9b.$$

The zero-th moment of the given ordinates is likewise the sum of the ordinates or

$$2+9+14=25.$$

Equating,

$$3a+9b=25,$$

which is one of the two equations (since there are two unknowns) desired.

It is left for the student to determine the first moments of the polynomial and of the given values, and show that

$$(a+b)+3(a+3b)+5(a+5b)=2+3\cdot 9+5\cdot 14,$$

or

$$9a+35b=99.$$

Solving the two equations simultaneously, we obtain $a=-\frac{2}{3}$ and $b=3$ and the desired equation becomes

$$y=-\tfrac{2}{3}+3x.$$

To appreciate the result, the three points and the line just obtained should be plotted. If we substitute 1, 3 and 5 in the equation just found, we obtain $y=2\frac{1}{3}$, $8\frac{1}{3}$ and $14\frac{1}{3}$ respectively. These values of y are called *graduated* values of 2, 9 and 14.

The *graduation* of a set of observations or values evidently depends very vitally upon the selection of the function to which they are graduated.

EXERCISES

In selecting the form of a polynomial to be fitted to a set of data, the data should be plotted as a check upon the nature of the curve to be fitted; the successive differences should also be computed, and the order of differences which proves most nearly to be constant should decide what degree the polynomial to be fitted should be assumed to have.

1. The pressure (P) of water in pounds per square inch at different depths (D), in units of 10 feet, was found by experiment to be as follows:

P,	8.66	17.32	25.99	34.65	43.31
D,	2	4	6	8	10
P,	51.98	60.64	69.31	77.97	86.63
D,	12	14	16	18	20

Determine a relation of the form $P=aD+b$.

2. In experiments to determine the effort necessary to raise different loads with a crane, the following values were obtained:

E,	60	70	79.5	89.5	99	108.5	118.25	128	138	147.5
R,	1080	1283	1483	1683	1883	2083	2283	2483	2683	2883

Determine a relation between the load R and the effort E of the form $R=aE+b$.

3. Same as Problem (2) but using the following data:

E,	14.2	26.6	38.1	50	59.1	72	81.8	91.5
R,	28	56	84	112	140	168	196	124

4. In an experiment with a Weston differential pulley block, the effort E in pounds required to raise a load W, in units of 10 pounds, was found to be as follows:

W,	1	2	3	4	5	6	7	8	9	10
E,	3.25	4.875	6.25	7.50	9	10.50	12.25	13.75	15	16.50

Determine the relation $E=aW+b$.

5. The readings of a standard gas-meter S, and those of a meter T being tested on the same line, were found to be:

S,	3000	3510	4022	4533
T,	0	500	1000	1500

Determine the relation $S=aT+b$. What are the meanings of a and b?

6. The following observations were made: where y denotes the melting point (C.) of an alloy of lead and zinc containing x per cent of lead.

x,	40	50	60	70	80	90
y,	186	205	226	250	276	304

Determine the relation $y=a+bx+cx^2$. Suggestion: Express x in terms of units of 10's.

7. The distance S in feet passed over by a falling body in t seconds was found by experiment to be:

S,	0	5	16	35	65
t,	0	.5	1	1.5	2

Determine the relation $S=at^2$. Suggestion: Fit the line $S=au$ where $u=t^2$. *Ans.* $S=16.1t^2$.

8. Same as Problem (7) but using the data:

S,	3.1	13	30.6	50.1	79.5	116.4
t,	.5	1	1.5	2	2.5	3

9. The following data give the velocities of water in the Mississippi river at various depths x for the point of observation chosen, the total depth being taken as unity.

y,	3.1950	3.2299	3.2532	3.2611	3.2516
x,	0	.1	.2	.3	.4
y,	3.2282	3.1807	3.1266	3.0594	2.9759
x,	.5	.6	.7	.8	.9

Ascertain by differences the appropriate degree for the polynomial $y=a+bx+cx^2+\ldots$ and determine the coefficients.

10. The pressure p measured in centimeters of mercury, and the

volume v, measured in cubic centimeters of a gas kept at a constant temperature, were found to be as follows:

v,	145	155	165	178	191
p,	117.2	109.4	102.4	95	88.6

Determine a relation of the form $pv=k$. Suggestion: Take the logarithm of both sides of the equation. (Express the data in logarithms.)

11. The amount of water A in cubic feet that will flow per minute through 100 feet of pipe of diameter d in inches, with an initial pressure of 50 pounds per square inch, was found to be as follows:

d,	1	1.5	2	3	4	6
A,	4.88	13.43	27.50	75.13	152.51	409.54

Find the relation $A=kd^n$. (Use logarithms). *Ans.* $A=4.88d^{2.473}$.

12. In testing a gas engine, corresponding values of the pressure p measured in lbs. per sq. ft. and the volume v in cubic ft. were obtained as follows:

v,	7.14	7.73	8.59
p,	54.6	50.7	45.9

Determine the relation $p=kv^n$. *Ans.* $p=387.6v^{-.938}$.

13. Same as Problem (12) but using the data:

v,	6.27	5.34	3.15
p,	20.54	25.79	54.25

Ans. $pv^{1.41}=273.5$.

14. Given the age and height in feet of a tree, as follows:

Age v,	13	34.4	50.5	218	247
x,	13.4	27.5	38.4	72.5	73

Determine the relation $v=kx^n$.

15. The specific gravity y of dilute sulphuric acid at different concentrations x per cent is given as follows:

x,	5	10	15	20	25	30	35
y,	1.033	1.068	1.101	1.139	1.178	1.218	1.257

Determine an appropriate relation.

49. Unit Moments: Quadrature Formulas.—It should be obvious that any particular moment of a distribution as heretofore defined may be made to assume any value we please by changing all the frequencies appropriately and proportionally; that is, the value of a particular moment depends upon the value of the total frequency. Unless, then, some standard for the total frequency or area is established, the values of the moments can not be controlled in a satisfactory manner. For this reason, a standard area or frequency is assumed. The most useful area to be taken as the standard is unity, and the moments computed on such a basis are called *unit* moments. The values of the unit moments could be obtained in the precise manner illustrated in the preceding section if each of the frequencies were first divided by the total frequency, but the same result can be obtained much more easily if the moments are first computed in the manner illustrated and then divided by the total frequency. It should be evident on a little investigation that the two plans must give identical results. If we let the symbols ν_0', ν_1', ν_2', etc., refer to the zero-th, first, second, etc., unit moments, then, for the distribution of litters of mice given above

$$\nu_0' = \frac{1}{1}\frac{2}{2}\frac{1}{1} = 1,$$

$$\nu_1' = \frac{5}{1}\frac{5}{2}\frac{5}{1} = 4.59,$$

$$\nu_2' = \frac{2}{1}\frac{9}{2}\frac{17}{1} = 24.1,$$

etc.

Particular attention is called to the fact that the first unit moment or

$$\nu_1' = \frac{\Sigma x f(x)}{N},$$

where $f(x)$ represents the frequency of the observation x and N is the total frequency, is the arithmetic average. The zero-th unit moment invariably has the value unity. Why?

As the only moments—with a few exceptions—to which we

shall refer in the future will be unit moments, we shall refer to them from now on by the single term " moment."

Just as the area of a rectangular histogram (or the sum of the ordinates) can rarely be expected to be more than an approximation of the area under the corresponding frequency curve, so the values of the moments (which include the area as a special case) of a rectangular histogram or frequency distribution can be expected to be no more than approximations of the corresponding moments of the corresponding frequency curve. It is highly desirable, therefore, to distinguish the moments of a *curve* by a suitable notation. Accordingly, we shall refer to the moments of a curve by the symbols μ_0', μ_1', μ_2', etc.; the notation for the corresponding moments of a frequency distribution (or ν_0', ν_1', ν_2', etc.) has already been given.

It is scarcely necessary to say that the values of the moments of a curve can be computed accurately only by means of the ordinary calculus, and that when we equate the corresponding moments of a curve and of a frequency distribution a discrepancy is thereby introduced. In most of the applications which we shall consider these discrepancies will not prove serious, but in applications where considerable refinement and accuracy are essential it would be practically necessary to apply certain corrections to the moments of the frequency distribution by means of formulas known as *quadrature* formulas. There are many forms of these quadrature formulas, but all are obtained on the assumption that the given frequencies have been fitted by a suitable curve; the quadrature formulas are then no more than formulas for the moments of these curves, or corrections to be applied to the moments (called the rough moments) of the given frequencies, expressed so that they can be readily adapted to any situation. No treatment of the method of moments as applied to statistical data can be regarded as complete without a complete treatment of quadrature formulas, but a complete treatment of such formulas would carry us so far from the main purposes of this course that it must be omitted. Nothing less than an intensive treatment would give the student a satis-

factory working knowledge of such formulas. Moreover, the choice of a suitable quadrature formula is very frequently open to controversy even among those who have considerable experience with work of this kind. It should be kept constantly in mind, however, from now on, that the moments of a frequency distribution are only approximations of the corresponding moments of the corresponding frequency curve.

50. Moments About the Mean.—We defined moments originally with respect to the y-axis; but it is obvious that the value of a moment will depend not only upon the value of the total frequency but also upon the position of the y-axis or the axis of reference. As the translation of the axes is a common procedure in mathematical analysis, it is only natural for purposes of clearness and definite understanding to establish a standard position for this axis of reference. The standard position of the axis of reference is taken at the mean, which is readily determined by computing the first moment. Reference to the method, given previously, of computing the correction to a provisional or trial mean shows that the value of the first moment about the mean is zero and that the mean is graphically the abscissa corresponding to the ordinate which bisects the total area, and, hence, which passes through the center of gravity of the area. The center of gravity is usually referred to in moments as the *centroid*, and the vertical axis of reference which passes through the centroid as the *centroid vertical*.

If we use the symbols adopted previously, but without primes, to refer to the moments about the mean, then, obviously

$$\nu_0 = \mu_0 = 1,$$

and

$$\nu_1 = \mu_1 = 0.$$

It is easy to determine the values of the moments about the mean from the values of the moments about any other axis of reference, by means of the lateral transformation used so frequently in elementary analysis.

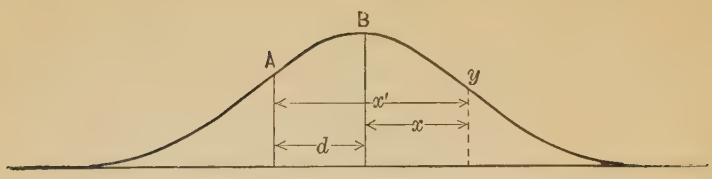


FIG. 2.

Let the distance between any axis of reference A about which the moments are known and the ordinate B at the mean be d. Then, if the distance between any ordinate y and A is x' and between y and B is x

$$x' = x + d,$$

and

$$(x')^n = (x+d)^n.$$

Hence, the n-th moment about the ordinate A is

$$\mu_n' = \frac{\Sigma(x')^n y}{N} = \frac{\Sigma(x+d)^n y}{N} \text{ (where } N \text{ refers to the total area)}$$

$$= \frac{1}{N}\left(\Sigma x^n y + nd\Sigma x^{n-1}y + \frac{n(n-1)}{2}d^2\Sigma x^{n-2}y + \text{etc.}\right)$$

$$= \mu_n + nd\mu_{n-1} + \frac{n(n-1)}{2}d^2\mu_{n-2} + \text{etc.},$$

or transposing

$$\mu_n = \mu_n' - nd\mu_{n-1} - \frac{n(n-1)}{2}d^2\mu_{n-2} - \text{etc.} \quad . \quad . \quad . \quad . \quad (41)$$

It should be noted that the formula just derived is an accumulative form; that is, moments about the mean are employed to find the next higher moment about the mean. It should be noticed also that the symbol μ would really be appropriate only if the summations considered above were performed by integration; since, however, formula (41) holds for either of the symbols μ or ν, the essential thing is to remember the distinction in the use of the symbols.

Substituting $n=0$ and $n=1$ in formula (41) we obtain

$$\mu_0=\mu_0'(=1),$$

$$\mu_1=\mu_1'-d\mu_0=\mu_1'-d.$$

But since

$$\mu_1=0,$$

$$\mu_1'-d=0,$$

or

$$d=\mu_1'. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (42)$$

The latter relation shows that the first moment about any ordinate (i.e., the arithmetic average of the deviations) is the directed distance from that ordinate to the mean. (Compare this fact with the method of computing the correction to be applied to the provisional mean considered in connection with the arithmetic average.) For example, suppose that the first moment of the frequency distribution of litters of mice were computed about the provisional mean $x=5$ as follows:

	y	x	xy
1	7	−4	−28
2	11	−3	−33
3	16	−2	−32
4	17	−1	−17
5	26	0	
6	31	1	31
7	11	2	22
8	1	3	3
9	1	4	4
	121		60
			−110
			− 50

Therefore

$$\nu_1'=\frac{-50}{121}=-0.41.$$

The mean is therefore the provisional mean plus the value of ν_1' or $5-0.41=4.59$, as found previously.

It is left for the student to obtain the following relations from formula (41):

$$\mu_2 = \mu_2' - d^2. \quad \text{Compare with (38).} \quad . \quad . \quad . \quad (43)$$

$$\mu_3 = \mu_3' - 3d\mu_2 - d^3. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (44)$$

etc.

EXERCISES

1–4. Compute the values of the first and second unit moments about the mean of the distributions given in the exercises following Art. 47.

5. The number of anal fin-rays was found, for 1000 minnows, to be:

Number	Frequency
13	5
12	144
11	554
10	279
9	15
8	2
7	1

Compute the first and second moments about the mean.

6. The weights of a large group of British males (adults) were found to be:

Weights (lbs.)	Frequencies	Weights (lbs.)	Frequencies
90	2	190	263
100	26	200	107
110	133	210	85
120	338	220	41
130	694	230	16
140	1240	240	11
150	1075	250	8
160	881	260	1
170	492	270	0
180	304	280	1
			7749

Compute the first and second moments about the mean. What is the mean?

7. Seed capsules of Shirley Poppies were found to have numbers of stigmatic rays in accordance with the following distribution:

Number of Rays	Number of Capsules	Number of Rays	Number of Capsules
6	3	14	302
7	11	15	234
8	38	16	128
9	106	17	50
10	152	18	19
11	238	19	3
12	305	20	1
13	315		

Compute the first and second moments about the mean. What is the mean?

8. The head-breadths of 1000 students at Cambridge University were measured, to the nearest one-tenth of an inch, to give the following distribution:

Head-breadth	Frequencies	Head-breadth	Frequencies
5.5	3	6.2	142
5.6	12	6.3	99
5.7	43	6.4	37
5.8	80	6.5	15
5.9	131	6.6	12
6.0	236	6.7	3
6.1	185	6.8	2
			1000

Compute the first and second moments about the mean.

51. The Standard Deviation.—The standard deviation has already been defined and is also evidently the square root of the second (unit) moment about the mean, or

$$\sigma = \sqrt{\mu_2} = \sqrt{\mu_2' - d^2}. \quad . \quad . \quad . \quad . \quad . \quad (45)$$

In the case of the frequency distribution of litters of mice,

$$\sigma = \sqrt{24.1 - (4.59)^2} = 1.75 \text{ approximately.}$$

It is left for the student to obtain the same value by taking the trial mean at $x = 4$ in the same distribution.

It is well to refer again to other names for the standard deviation, namely: dispersion, root-mean-square, mean error, etc. It is well also to refer again, for purposes of emphasis, to the great importance of the standard deviation in weighing the relative consistencies of sets of deviations. This use of the standard deviation or dispersion will form the basis of practically all of the work of the succeeding chapters.

EXERCISES

Compute the value of the standard deviation of the distributions given in the preceding list of exercises. Remember to take the original unit of measurement into consideration.

52. Secular Trend of a Time Series.—If we plot the terms of a time series—say by months—for a considerable number of years in the customary way, the vagaries of the graph are not apt to impress the novice with any possibilities of forecasting general business conditions because he will at least know that there are many factors affecting the graph. It has been found possible, however, to classify and then to eliminate these effects down to those that are more directly related to the recurrent periods of prosperity and depression. We shall refer to these classes of variations as *secular trend*, *seasonal variation* and *cyclical fluctuation* (including irregular variations).

We shall consider first secular trend or that part of the fluctuation which tends to extend consistently throughout the whole period under consideration. The standard method of treating secular trend is to fit a curve—usually a straight line—

to the totals given by years. We shall consider only those series of totals whose graphs justify the fitting of a straight line. In the light of the preceding work of this chapter such a fitting involves no difficulties whatever, but it is well to state that the origin in all such examples is taken at the middle of the given period and that the graduated values must correspond to the middle of each year.

Each fitting could, of course, be made independently but something will be gained if we derive the fundamental equation directly in terms of the proper moments once for all by fitting the equation $y=mx+b$ to the pairs of values (x_1, y_1), (x_2, y_2), . . . (x_n, y_n).

For the zero-th moment we have

$$\Sigma(mx+b)=m\Sigma x+nb=\Sigma y,$$

and for the first moment

$$\Sigma x(mx+b)=m\Sigma x^2+b\Sigma x=\Sigma xy.$$

Since, however, by hypothesis $\Sigma x=0$ these equations reduce to

$$nb=\Sigma y,$$

and

$$m\Sigma x^2=\Sigma xy.$$

Our final equation then becomes

$$y=\frac{\Sigma xy}{\Sigma x^2}x+M_y, \quad . \quad . \quad . \quad . \quad . \quad . \quad (46)$$

where M_y is the mean of the y's.

The expression Σxy refers to a new kind of moment of the second order—known as the *product moment*—which will receive complete treatment in a later chapter, but its meaning is so clear that no further discussion is necessary here. The

application of equation (46) to the production of pig iron for the period 1903–16 is shown below.

Year	Pig Iron (1000 long tons)	Deviations y	$2x$	$2xy$	$4x^2$
1903	17,426	− 7,521	−13	97,773	169
1904	16,131	− 8,816	−11	96,976	121
1905	22,587	− 2,360	− 9	21,240	81
1906	24,791	− 156	− 7	1,092	49
1907	25,307	360	− 5	− 1,800	25
1908	15,620	− 9,327	− 3	27,981	9
1909	25,397	450	− 1	− 450	1
1910	26,849	1,902	1	1,902	1
1911	23,326	− 1,621	3	− 4,863	9
1912	29,375	4,428	5	22,140	25
1913	30,716	5,769	7	40,383	49
1914	23,050	− 1,897	9	− 17,073	81
1915	29,661	4,714	11	51,854	121
1916	39,025	14,078	13	183,014	169
	14)349,261			2)520,170	4)910
	24,947			260,085	227.5

The equation of the trend line is then $y=\frac{260085}{227.5}x+24947$ or $y=1143x+24947$ (annual production). If we divide this equation through by 12 we obtain a more useful equation—the equation giving the monthly production

$$y=95.3x+2079,$$

where, however, x is still measured in years.

It is now easy to obtain the graduated series. The work can be carried out very effectively by using the latter equation for, say January of each year (*but for the middle of the month*) and then adding the monthly difference of monthly production to obtain the productions for all the other months. Thus, the production for January, 1910, would be $95.3(\frac{1}{12})+2079$ or

2083; and the production for January, say 1913, would be $95.3(\frac{3}{12}\frac{7}{})+2079$ or 2373. To obtain the production for February, 1910 (or 1913), we add the monthly difference 95.3 to 2083 (or 2373). (What would constitute a good check upon the work?)

We are interested primarily in the *elimination* of the secular trend from the original series *by months* and this is obtained by taking *percent relatives of the original data to their respective graduated values*. The method should be clear after verifying the statement that the percent relative for January 1910 is $100\frac{2608}{2083}$ or 125% where the production given originally for that month is given to be 2608 (see the table given in Ex. 1 following Art. 54); and so on for the rest.

It would be very desirable and even necessary in practice to plot the graph of the values of the relatives just found but a detailed discussion of that important but elementary process is unnecessary here. However, the graph is given at the end of this chapter (Chart II, Art. 54) and merits considerable study. One will find that the trend is now horizontal or that the secular trend which we have defined more or less by the method of attack has disappeared. We shall consider in the next section the measure and the elimination of the seasonal variation.

53. Seasonal Variation.—It is scarcely necessary to define what we mean by seasonal variation since every one is familiar with the fact that prices, production, etc., have a tendency—more in some cases than in others—to be greater or less in some months than in others.

In analyzing seasonal variation we shall employ what is known as the *median-link relative* method.[1] The first step consists in setting up the link relatives (see Art. 46) of a time series by months. Such a table is as follows:

[1] See Person's article in Rev. Econ. Stat., Prel., Vol. I, Jan., 1919, for derivation and detailed discussion.

LINK RELATIVES FOR THE MONTHLY TONNAGE PRODUCTION OF PIG IRON IN THE UNITED STATES

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1903	....	94	114	101	106	98	92	102	99	92	73	81
1904	109	131	120	107	99	84	86	106	116	107	102	109
1905	110	90	121	99	102	91	97	106	103	108	98	102
1906	101	92	113	96	101	94	102	96	102	112	100	102
1907	99	92	109	100	104	97	101	100	97	107	78	67
1908	85	103	114	94	101	94	112	112	104	111	101	110
1909	103	95	107	95	108	103	109	107	106	109	98	103
1910	99	92	109	95	96	95	95	98	98	102	91	93
1911	99	101	122	94	92	95	100	107	103	106	95	102
1912	101	102	114	99	106	97	99	104	98	109	98	106
1913	100	92	107	100	103	93	97	100	99	102	88	89
1914	95	100	124	97	92	92	102	102	94	94	85	100
1915	106	105	123	102	107	105	108	108	103	110	97	105
1916	100	97	108	97	104	96	100	99	100	110	94	96
1917	99											

We are interested, however, how, *on the average*, the value of each month is related to the value for the preceding month of the year. We are interested then in some form of an average of the link relatives of each month but for all the years under consideration. In a problem like this the number of years and, therefore, the number of variates making up any average is almost always small. The arithmetic average is therefore likely to prove unreliable due to the presence of extreme variations; for this reason the median is apt to prove preferable and in accordance with the method adopted above we shall employ the median, although it would be desirable in any important investigation to utilize some form of graph whereby the eye may be utilized to estimate the most representative value for this average and to hold little faith in any such value, however selected, unless there is considerable consistence between all the values. One method of dealing with cases in which the average is not clearly indicated would be to

take the arithmetic average of the middle four or of the middle three values. The medians for the table given above will be found to be

January	100	May	102.5	September	101
February	96	June	95	October	107.5
March	114	July	100	November	96
April	98	August	103	December	102

The next step is to change our representative link relatives to fixed base relatives and this may be performed by forming chain relatives. Since the chain relative of any month is the product of all the preceding links it is well to resort to logarithms. The link relatives given just above and their logarithms (omitting -10 from any characteristic) are given below with other columns of values to be explained.

Analysis of Seasonal Variation for Pig Iron Production for 1903–16

	Link Relatives	Logs	Logs -0.0049	Logs of Chain Relatives	Chain Relatives	Adjusted Chain Relatives
January	1.00	0.0000	9.9951	9.9951	98.9	99
February	.96	9.9823	9.9774	9.9725	93.9	94
March	1.14	.0569	.0520	.0245	105.8	106
April	.98	9.9912	9.9863	.0108	102.5	103
May	1.025	.0107	.0058	.0166	103.9	104
June	.95	9.9777	9.9728	9.9894	97.6	98
July	1.00	.0000	9.9951	9.9845	96.5	97
August	1.03	.0128	.0079	9.9924	98.3	98
September	1.01	.0043	9.9994	9.9918	98.1	98
October	1.075	.0314	.0265	.0183	104.3	104
November	.96	9.9823	9.9774	9.9957	99.0	99
December	1.02	.0086	.0037	9.9994	99.9	100
		12) .0582			1198.7	
		.0049				

The chain relatives to be computed as outlined above will necessarily have the preceding December as a base. This base should be changed to one corresponding to the current year. Let us note first, however, that if the seasonal effect were the only cause of variation and the median link relatives were truly representative the chain relative for December based upon the preceding December would have to be 100%. The logarithm of the actual chain relative is, however, 0.0582 and it is only reasonable to expect better results if we make it what it should be by deducting one-twelfth of 0.0582 from each of the logarithms of the link relatives before we proceed. The next column then gives the logarithms of the chain relatives or the sum of the corrected logarithms to the left beginning with the one immediately to the left and summing upward. The column which is the last but one gives the chain relatives themselves. Finally, the base is changed from the preceding December to the (average) current year by dividing each of these chain relatives by one-twelfth of their sum to give our final estimates of the seasonal variation which appear in the last column—rounded off to the nearest per cent.

To *eliminate* seasonal variation as defined more or less by the method of attack we *subtract the values of the seasonal variation* (as obtained above) *from the terms of the series obtained in eliminating secular trend*—that is, from the per-cent ratios of the original items to the graduated items. The scheme is based upon the assumption that the seasonal variation is a fairly uniform percentage of the normal trend—a rather questionable assumption but the best that has been offered—and the subtraction of the estimates of the variation from the ratio of the actual item to the graduated item really amounts to subtracting from the actual item a percentage of the graduated item and dividing the resulting deviation by the graduated item.

Here, again, the plotting of the graph of the items of a time series with both secular trend and seasonal variation

eliminated (see charts at the end of the next Art.) would be all important in practice and in this case would lead to especially important conclusions. A discussion of these conclusions will be given in the next section.

54. Cyclical Fluctuations. Lag.—The graph mentioned at the last of the preceding section is of the greatest importance in forecasting general business conditions because it reflects approximately only those fluctuations which are usually associated with the idea of general prosperity or of general depression *except* possibly exceptional fluctuations due to factors—such as war, panics, etc.—which are so irregular and infrequent that we can never hope to predict or control them from any statistical knowledge of their past behavior. On the other hand, whenever an irregular factor of unusual importance is known to have wielded a great influence during a certain period it should be obvious that any graph of a time series for that period is apt to prove deceptive if not worthless as a basis for forecasting periods of general prosperity or depression. In any case we shall refer to all fluctuations in a time series in which secular trend and seasonal variation have been eliminated as *cyclical fluctuations* and hope for the absence of the effects of disturbing irregular factors.

The graph of cyclical fluctuations might well be regarded as our goal if the basic series were our whole concern and the recurrences of the peaks and dips of the graph, however irregular, would form an admirable basis for study and even for a fair amount of forecasting of conditions connected with the commodity or item under investigation. Such a study would reveal information in regard to the recurrence of particularly favorable or unfavorable periods; it would reveal also, however, that these recurrences are not sufficiently regular to warrant predictions by mathematical processes. But it will be found that the graphs of the cyclical fluctuations of several commodities or items will differ widely because of differing degrees of dispersion from the normal trend and for purposes of comparison it is practically necessary

to express all terms in units of the standard deviation. This plan is open to some criticism in dealing with a time series but it is the best at our disposal and is easily carried out because the data used to obtain the graph are already in an approximate form of deviations from the mean and it is simply necessary to square each item, compute the mean of the squares and extract the square root of this mean. Each item of the data is then divided by this value of the dispersion to form what are called *cycles*.

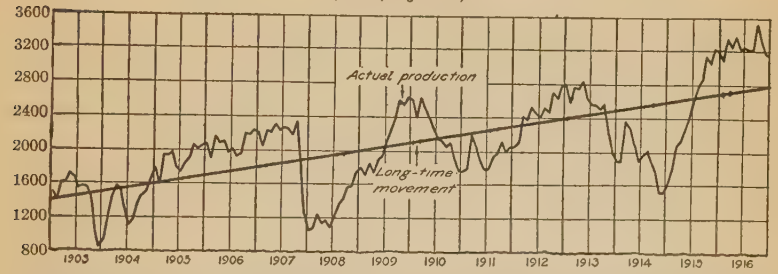
The graphs of the cycles of any two series can now be compared with considerable ease and assurance. Some of the most interesting comparisons of this kind have been made between production series of certain important commodities and series of interest rates, for in some cases the graphs are very similar except in one important particular: corresponding points on the two graphs do not correspond to the same time and this difference in time is usually fairly constant throughout the graphs. This very common phenomenon of economic cycles is known as *lagging correlation* and the length of time by which one curve precedes the other is called the *lag*. Although efforts have been made to establish methods[2] for making accurate measurements of lag the ordinary rules for numerical computation seem to warrant the assumption of little more accuracy than that with which the lag can be estimated simply from the inspection of the graphs. We may even go farther and say once for all that all of the methods used in treating time series which we have considered in the last few sections can be expected to lead only to general conclusions to be checked and rechecked by numerous investigations. The inventors of these methods have been wise in admitting this fact and in placing the greatest emphasis upon the necessity

[2] One method is to compute the sum of the products of the corresponding cycles of the two series for several trial (integral) values of the lag; the best trial value of the lag will give the greatest sum of products because such a sum is the value of the correlation coefficient (treated in Chapter X).

of the investigator always keeping in close touch with his data and any apparent discrepancies. The following charts illustrate the work of this and the immediately preceding articles.

CHART I. — PIG-IRON PRODUCTION: ACTUAL FIGURES

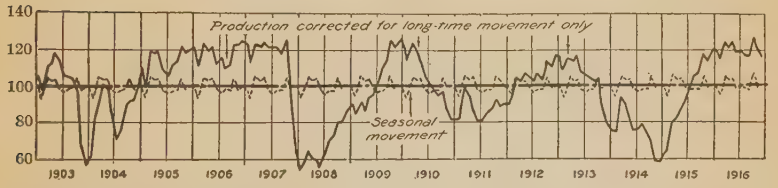
(*Unit: 1,000 gross tons*)



EXPLANATION: The above chart shows the monthly production of pig iron for the years 1903–16. As the United States, throughout this period, grew in population and in the extent of its commercial activities, new iron furnaces were set in operation and there was a gradual expansion in the total output of the nation. This growth element, or long-time movement upward, in the normal producing capacity of this industry has been represented in the chart by the straight line fitted to the original items. Correction for this influence was accomplished by expressing the actual tonnage manufactured in each month as a percentage of the corresponding "normal" production indicated by the straight line. The resulting figures appear in Chart II below

CHART II. — PIG-IRON PRODUCTION: FIGURES CORRECTED FOR LONG-TIME MOVEMENT

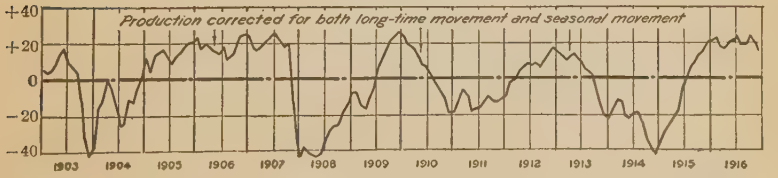
(*Percentages*)



EXPLANATION: The solid line on the above chart represents pig-iron production corrected for the long-time movement, but not for the seasonal movement; the dotted line indicates the usual seasonal variation in the output during the year. Correction for the seasonal influence was made by taking the differences between the corresponding items of these two lines; the results of this subtraction show in each case by how much the actual production, after the long-time movement had been eliminated, exceeds or falls short of the output normally to be expected for that month. These differences (presented in Chart III below) are called "the corrected figures" since both the long-time and seasonal movements have been removed.

CHART III. — PIG-IRON PRODUCTION: CORRECTED FIGURES

(*Percentages*)



Courtesy of Harvard Economic Service

EXERCISES

1. The part of the time series of production of pig iron used in Art. 52 which corresponds to the months of the year 1910 is as follows:

January.....	2608	May.....	2390	September.....	2056
February.....	2397	June.....	2265	October.....	2093
March.....	2617	July.....	2148	November.....	1909
April.....	2483	August.....	2106	December.....	1777

(*a*) Compute the terms of the graduated series which correspond to these values.

(*b*) Determine the terms to be obtained by eliminating the secular trend.

(*c*) Plot the (two) graphs of the terms found in (*a*) and (*b*).

2. Determine the equation of the trend line for monthly production for the data giving the production of pig iron given in Art. 52 but with the production 38,185 for 1917 added. *Ans.* $y=105x+2152$.

Given the monthly bank clearings in New York City in units of $10,000,000 as follows:

	Jan.	Feb.	March	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1903	719	524	600	594	557	575	598	466	429	523	461	550
1904	569	443	480	479	484	457	519	466	535	748	834	850
1905	773	712	873	868	788	674	684	698	686	803	854	969
1906	1124	832	838	854	879	782	726	883	831	934	861	923
1907	964	733	956	767	733	637	731	689	603	820	550	535
1908	675	490	550	549	667	565	662	623	678	723	819	927
1909	888	689	749	861	810	911	832	869	848	1005	906	991
1910	1125	815	905	834	781	836	788	646	623	790	781	802
1911	853	730	762	697	796	822	752	738	718	751	807	811
1912	884	731	842	883	878	796	792	749	743	1014	882	881
1913	934	779	789	806	796	775	734	676	746	869	759	800
1914	937	724	785	855	724	784	818	458	463	561	540	653
1915	729	648	757	881	864	803	870	854	962	1274	1183	1233

3. Plot the graph of the data.

4. Find the series to be obtained by eliminating the secular trend and plot its graph.

5. Eliminate the seasonal variation and plot the resulting graph (with secular trend also eliminated).

The following table gives the per cent deviations from secular trend corrected for seasonal variation for pig iron production for the specified years.

	Jan.	Feb.	March	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1903	55	4	5	9	14	17	9	8	7	− 8	−30	−44
1904	−38	−15	−11	− 1	− 5	−14	−26	−24	−12	−12	− 6	1
1905	12	5	13	15	16	11	9	13	15	19	20	21
1906	23	17	20	18	17	16	18	11	13	20	24	25
1907	24	19	17	19	21	24	26	23	19	21	− 2	−34
1908	−44	−37	−42	−43	−44	−41	−34	−29	−26	−24	−19	−12
1909	− 8	− 8	−14	−16	−11	− 3	7	12	18	23	24	27
1910	26	20	19	15	9	9	4	0	− 3	− 6	−11	−18
1911	−18	−12	− 6	− 9	−18	−17	−16	−12	−10	−10	−11	−10
1912	− 8	− 2	− 1	1	5	8	7	9	7	11	13	18
1913	19	15	10	12	14	11	9	6	5	1	− 8	−19
1914	−22	−18	−11	−11	−20	−21	−19	−19	−24	−33	−40	−41
1915	−36	−29	−25	−21	−17	− 6	2	8	10	15	16	21
1916	21	22	19	18	21	21	23	20	20	25	22	16

6. Show that the mean of all these values is a little over one tenth of one unit.

7. Show that the value of the standard deviation is about 19.1.

8. Form a table of cycles of these values.

CHAPTER VIII

THE NORMAL CURVE

55. The Normal Curve.—We referred in earlier sections to the tendency of a certain kind of errors to compensate or offset each other. It will be remembered also that we called attention to the fact that certain kinds of "deviations," "residuals," etc., behaved in like manner and that we would agree to refer to the whole general class of such items by the term "errors." Suppose that we have a large number, say several thousand, of such errors which compensate in the most ideal manner, and suppose that these errors are recorded in classes according to sizes, to give corresponding frequencies. The question naturally arises: What would be the most natural form of the corresponding frequency curve under such ideal conditions? We should probably all agree upon the two most important characteristics of the curve. We should expect a considerable "hump" or a point of maximum at the mean—that is, we should expect the smallest errors to occur most frequently—and we should expect the curve to approach the x-axis on both sides of the vertical axis of reference taken at the mean in the same way, so that the curve would be symmetrical with respect to this axis. Such a frequency curve would look much like one of the following curves and is known by several names: the normal curve, the probability curve, the error curve, the Gaussian [1] curve, etc. We shall refer to it as the normal curve, and to any frequency distri-

[1] As a matter of fact, Laplace was an earlier contributor of knowledge concerning the curve than Gauss.

bution whose corresponding frequency curve is a normal curve as a normal distribution.

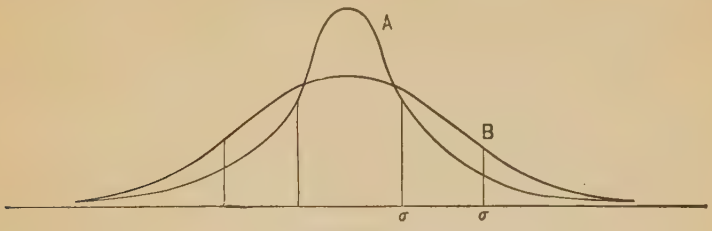


FIG. 3.

Now it is only too obvious that we can no more expect a frequency distribution, even under the most ideal conditions, to prove truly normal in practice than we can expect an *a priori* probability to be verified exactly in a large number of trials in practice. A certain amount of variation is to be expected under any circumstances, and the discrepancies will vary in different distributions all the way from discrepancies which are so small that they would naturally be expected to discrepancies large enough to rule the given distribution out of further consideration as a normal distribution for practical purposes. Since truly normal distributions can not be expected in practice, we shall take the liberty of referring to many frequency distributions as normal distributions which are only approximately so but which differ from truly normal distributions so little that the variation might easily be ascribed to random variation. Thus, the following distribution, which gives the actual results of shooting one thousand times at a target consisting of a vertical line, would probably be regarded as a normal distribution in the sense expressed above:

Deviations x	Frequencies y
−5	1
−4	4
−3	10
−2	89
−1	190
0	212
1	204
2	193
3	79
4	16
5	2
	1000

After all, whether a given distribution is normal or not normal is not of great importance. We shall find as we proceed that the matter which is of the greatest importance is whether the distribution, of which we may have only a few sample numerical observations, could be assumed to be sufficiently normal to permit us to draw certain conclusions which we shall consider later. There are diversities of opinion and attitude in regard to the whole question, and no absolute criterion or rule can be laid down; but it has been found from experience that the vast majority of all frequency distributions of deviations from the mean of numerical observations made upon a size or characteristic *of a single object*, where there is cogent reason for believing that deviations in one direction are just as probable as deviations in the other direction, are normal distributions in the sense adopted above. On the other hand, distributions of deviations of observations made upon a characteristic *of several objects*, even though those objects belong to the same general class, can rarely be expected to be normal. Thus, the distribution of a large number of refined readings of a barometer, made under proper conditions, is very apt to be normal, while the distributions of lengths of a large number of ears of corn or of leaves of trees, of the various incomes or of popu-

lations by ages of a large community, are apt to vary considerably from normal distributions.

56. The Derivation of the Equation of the Normal Curve.— If we take the origin at the mean of the normal curve, the slope of the curve at $x=0$ must be zero. Moreover, the curve would approach the x-axis alike on both sides of the y-axis so gradually that the slope of the curve would approach zero as x increased in absolute value without limit or, what amounts in this case to the same thing, as y approached zero. All these properties of the curve can be expressed algebraically by the differential equation

$$\frac{dy}{dx}=-kxy,$$

where the negative sign is inserted to insure that there shall be a maximum and not a minimum at $x=0$, and k is a constant. As a matter of fact, it does not follow at all from what has been said that k must be a constant; but if the slope of the curve is to be zero nowhere else than where we have indicated, k must be either a constant or a function of the form $1/f(x)$. It remains, then, merely to explain that the differential equation corresponding to the latter alternative has been investigated at great length, especially for the case where it is assumed that $f(x)$ can be expanded in the form of a power series. The latter assumption leads to a wonderful system of frequency curves (known as the Pearson types[2] of frequency curves) which include the normal curve as a special case, where k is a constant. It is an interesting fact in that connection that practically none of these frequency curves (except the normal curve) is symmetrical with respect to the vertical axis of reference and some of them do not have a maximum at $x=0$ or have no maximum at all. (How can these facts be reconciled with the form of the differential equation?) We are justified, then, in assuming k to be a constant.

[2] See Pearson's "Tables for Statisticians and Biometricians."

Writing the differential equation given above in the form

$$\frac{1}{y}\frac{dy}{dx} = -kx,$$

and integrating, we obtain

$$\log y = -k\frac{x^2}{2} + \log y_0,$$

where the constant of integration is written as a logarithm for purposes of combination. The final equation can then be written

$$y = y_0 e^{-k\frac{x^2}{2}}.$$

It simply remains, then, to investigate the constants k and y_0 and show that they can be expressed in terms of characters with which we are already familiar. If we let N denote the area under the curve then

$$N = \int_{-\infty}^{\infty} y_0 e^{-k\frac{x^2}{2}} dx.$$

If this integration be performed by parts, with $dv = dx$ and u as the exponential expression, we obtain

$$y_0 \int_{-\infty}^{\infty} e^{-k\frac{x^2}{2}} dx = y_0 \left(x e^{-k\frac{x^2}{2}} \right)_{-\infty}^{\infty} + y_0 k \int_{-\infty}^{\infty} x^2 e^{-k\frac{x^2}{2}} dx.$$

The first expression on the right evidently takes the indeterminate form $\frac{\infty}{\infty}$ for each limit but is found, by following the process outlined in a preceding section, to have the value zero. The integral on the right is to be recognized as Nk times the second unit moment μ_2 of the normal curve. Hence, we have

$$N = Nk\mu_2,$$

or

$$k = \frac{1}{\mu_2} = \frac{1}{\sigma^2},$$

where σ is the standard deviation of the curve.

The definite integral given above can be valuated in another way. Referring to equation (34) in the chapter on Gamma

and Beta functions, it should be evident that the value of the integral is also

$$N = y_0\sqrt{\frac{2\pi}{k}} = y_0\sigma\sqrt{2\pi}.$$

Hence

$$y_0 = \frac{N}{\sigma\sqrt{2\pi}}.$$

The equation of the normal curve may then be written in the final form

$$y = \frac{N}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (47)$$

It will be recalled that points of inflection of a curve are points where the curve leaves off being concave downwards to become concave upwards, or vice versa, and that such points are found by equating the second derivative to zero and solving. It is easily verified by that process that the points of inflection of the normal curve are at $x = -\sigma$ and $x = \sigma$.

The equation of the normal curve is sometimes expressed in terms of what is called the *modulus* c, which is defined by the relation $c = \sigma\sqrt{2}$, and sometimes in terms of what is called the *precision* h, which is defined as the reciprocal of the modulus or $h = 1/c$.

57. Graduations of Normal Distributions: Tables of Ordinates.—The final form of the equation of the normal curve derived in the preceding section involves N and σ (and the constant π); the former is the area under the curve and would be given approximately by the total frequency of the corresponding distribution; an approximate value of σ would be given by the value of the standard deviation of the distribution. As the points of inflection of the curve are located at $x = \sigma$ and $x = -\sigma$, it is evident that for large values of σ the curve is dispersed or spread out somewhat like curve B (in Art. 55) and that for small values of σ the curve assumes a more compressed or steeper form similar to curve A. The *form* of the curve then depends solely upon the value of σ (assuming the units of measurement upon the two axes to be the same); in

fact, if we replace x/σ by x in the equation and $y\sigma/N$ by z, we obtain

$$z = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

which is obviously independent of σ. Extensive tables[2] of values of z have been compiled which are to be entered with values of x/σ to give values of z which may be adjusted to correspond to the frequencies of any normal distribution; a very small table is given at the end of the next article.

Given any normal distribution, suppose that the values of the mean and the standard deviation σ are computed; and then that a column of values of x/σ corresponding to the values of x with the origin taken at the mean is set up, and these values of x/σ are used to enter a table of values of z. If the values of z so obtained are then finally multiplied by N/σ, we obtain new or theoretical values of y. This column of theoretical values of y constitutes what is called a *graduation* or "fit" of the observed frequencies given originally. It should be obvious that a column of values of z/σ constitutes a column of probabilities of the occurrences of the corresponding deviations. This explains why the normal curve is frequently referred to as the probability curve. As an example, let us graduate the following distribution of certain measurements (cephalic indices) of Bavarian skulls. It is easily verified that the mean is 83.148 and that the standard deviation is 3.32. Hence, $N/\sigma = 900/3.32 = 271$. As the short table of ordinates given in the next section is used, no great refinement in computation will be employed; for example, the frequencies of "75 and under" and "92 and over" are treated as single ordinates in computing the mean and standard deviation without serious effects, but the theoretical frequencies for those intervals are determined by classes and then combined. The details of the process of graduation should be clear after a little study of the following results.

[2] Pearson's " Tables."

Measurements, x' (m.m.)	Frequencies, y'	Deviations, x	x/σ	z	$y=z\frac{N}{\sigma}$ (to the nearest integer)	
		10.148	3.06	0.00370	1.0	
		9.148	2.75	.00909	2.5	
75 and under	9.5	8.148	2.45	.01984	5.4	9
76	12.5	7.148	2.15	.03955	....	11
77	17	6.148	1.85	.07206	....	20
78	37	5.148	1.55	.12001	....	33
79	55	4.148	1.25	.18265	....	50
80	71.5	3.148	0.948	.25840	....	70
81	82	2.148	0.647	.32652	....	88
82	116	1.148	0.346	.37732	....	102
83	98	0.148	0.0446	.39872	....	108
84	107	0.852	0.256	.38726	....	105
85	82	1.852	0.558	.34446	....	93
86	74	2.852	0.859	.28011	....	76
87	58	3.852	1.16	.20357	....	55
88	34.5	4.852	1.46	.13742	....	37
89	19	5.852	1.76	.08478	....	24
90	10	6.852	2.06	.04780	....	13
91	8	7.852	2.37	.02406	....	7
92 and over	9	8.852	2.67	.01130	3.0	5
		9.852	2.97	.00485	1.3	
	$N=900$	10.852	3.27	.00190	0.5	

58. Graduations of Normal Distributions: Tables of Areas. —A normal distribution can be graduated in another way—by means of a table of areas of the normal curve. If we denote the area under a normal curve of unit total frequency from $x=-x/\sigma$ to $x=x/\sigma$ by α, then

$$\tfrac{1}{2}(1+\alpha)=\int_{-\infty}^{x/\sigma}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}dx.$$

If we replace x/σ by x we obtain

$$\tfrac{1}{2}(1+\alpha)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{x^2}{2}}dx.$$

It is easily seen that since $\frac{1}{2}(1+\alpha)$ denotes the area under the normal curve from the extreme left to the positive abscissa x/σ, then $\frac{1}{2}(1-\alpha)$ denotes the area from $x=x/\sigma$ to the extreme right. Extensive tables[3] of values of $\frac{1}{2}(1+\alpha)$ have been compiled which are to be entered with values of x/σ; a small table is given at the end of this article.

In graduating a normal distribution by means of a table of areas, class marks must be replaced by class limits—the class limits on the farther side from the mean. After the values taken from a table of areas have been multiplied by the total frequency N, each area so obtained must be deducted from the next greater area to obtain the successive class frequencies, *except* for the one class which contains the mean, in which case the areas on the two sides of the mean must be added. The process should be clear after a study of the following graduation of the distribution of 1000 shots given at the beginning of this chapter (Art. 55). It is easily verified

Frequencies, y'	Class Marks	Class Limits, x	$\frac{x}{\sigma}$	$\frac{N}{2}(1+\alpha)$	Theoretical Frequencies, y
1	5.48	5.98	3.80	1000	1
4	4.48	4.98	3.16	999	5
10	3.48	3.98	2.52	994	23
89	2.48	2.98	1.90	971	77
190	1.48	1.98	1.25	894	162
212	0.48	0.98 0.02	0.62 0.01	732 504	236
204	0.52	1.02	0.65	742	236
193	1.52	2.02	1.28	900	158
79	2.52	3.02	1.91	972	72
16	3.52	4.02	2.54	994	22
2	4.52	5.02	3.18	999	5
					1
1000					1000

[3] Pearson's "Tables."

that the mean is 0.48 and $\sigma = 1.58$; $N = 1000$. The second column gives the class marks measured from the mean and the third column gives the corresponding class limits to be used to enter the table of areas (the small table given at the end of this article is used).

It should be noted that in the case of the class frequency 236 the frequency 232 ($= 732 - 500$) is on one side of the mean and the frequency $4 (= 504 - 500)$ is on the other.

It need scarcely be stated that graduations to the normal curve can not be expected to be very satisfactory unless the given distribution is itself normal; otherwise, all that can be

TABLES OF ORDINATES AND AREAS OF THE NORMAL CURVE

x/σ	Ordinates, z	Areas, $\frac{1}{2}(1+\alpha)$	x/σ	Ordinates, z	Areas, $\frac{1}{2}(1+\alpha)$
0.0	0.39894	0.50000	2.0	0.05399	0.97725
0.1	.39695	.53983	2.1	.04398	.98214
0.2	.39104	.57926	2.2	.03547	.98610
0.3	.38139	.61791	2.3	.02833	.98928
0.4	.36827	.65542	2.4	.02239	.99180
0.5	.35207	.69146	2.5	.01753	.99379
0.6	.33322	.72575	2.6	.01358	.99534
0.7	.31225	.75804	2.7	.01042	.99653
0.8	.28969	.78814	2.8	.00792	.99744
0.9	.26609	.81594	2.9	.00595	.99813
1.0	.24197	.84134	3.0	.00443	.99865
1.1	.21785	.86433	3.1	.00327	.99903
1.2	.19419	.88493	3.2	.00238	.99931
1.3	.17137	.90320	3.3	.00172	.99952
1.4	.14973	.91924	3.4	.00123	.99966
1.5	.12952	.93319	3.5	.00087	.99977
1.6	.11092	.94520	3.6	.00061	.99984
1.7	.09405	.95543	3.7	.00042	.99989
1.8	.07895	.96407	3.8	.00029	.99993
1.9	.06562	.97128	3.9	.00020	.99995
			4.0	.00013	.99997
			4.1	.00009	.99998
			4.2	.00006	.99999

said of the graduation is that it constitutes the best fit *to the normal curve* that is possible; a graduation to some other curve would in that case probably prove more satisfactory. If the given distribution is normal it is reasonable to expect a graduation to the normal curve to be the best fit regardless of the curve employed. However, though the normal curve is clearly the curve to which a given distribution should be graduated, the given frequencies may be so "rough"—probably because of the relatively small number of observations—that the graduated or theoretical frequencies may not fit the observed frequencies very closely. In such a case the graduation must necessarily be a poor one but through no fault of the method of graduation.

EXERCISES IN GRADUATION

The following distributions give (1) the statures of 802 Cairo-born Egyptians and (2) the statures of 739 Smith College girls (1914–15). Fit each to the normal curve (*a*) using the table of ordinates, and (*b*) using the table of areas.

Cms.	(1)	Inches	(2)
149.5	4	59	1
153.5	28	60	2
157.5	73	61	2
161.5	185	62	11
165.5	212	63	11
169.5	167	64	48
173.5	87	65	45
177.5	31	66	97
181.5	14	67	100
185.5	1	68	126
		69	103
		70	97
		71	45
		72	46
		73	4
		74	1

EXERCISES

It has been found possible to select the most appropriate type of the Pearson frequency curves for fitting a given distribution from values of the following functions of moments:

$$\beta_1=\frac{\mu_3^2}{\mu_2^3},$$

$$\beta_2=\frac{\mu_4}{\mu_2^2},$$

computed for the given distribution.

1. Show by actual integration that for the normal curve $\beta_2=3$. (Suggestion: Integrate the expression for μ_2 by parts with $u=x^3$.)

2. Show by integration that for the normal curve $\beta_1=0$.

3. Show that $\beta_2=3$ for the normal curve, by integrating the expression for μ_2 by parts with $dv=x^2dx$.

4. Show that for the normal curve and for n, an even number,

$$\mu_n=\frac{\mu_{n+2}\cdot}{(n+1)\sigma^2}.$$

Integrate μ_n by parts with $dv=x^ndx$.

5. Show that for the normal curve

$$\frac{\mu_6}{\mu_2\mu_4}=5,$$

and

$$\frac{\mu_8}{\mu_2\mu_6}=7.$$

6. Compute the value of β_2 for (*a*) the distribution of shots and (*b*) the distribution of measurements of Bavarian skulls, given in the text.

7. Assuming the distributions graduated in the text to be truly representative, determine the values of the probabilities:

(*a*) Of the occurrence of a cephalic index of 80 mm.

(*b*) Of the occurrence of a cephalic index greater than 80 mm.

8. Referring to the tables of ordinates and areas, what are the probabilities:

(a) Of an occurrence in a normal distribution of a deviation of 1.46σ? Of a deviation of 3.84σ?

(b) Of an occurrence of a positive deviation greater than 1.46σ? Of any deviation greater than 3.84σ?

(c) Show that the probability of obtaining any deviation greater than $\frac{x}{\sigma}$ (i.e., in absolute value) is $1-\alpha$.

(d) How could values of $1-\alpha$ be obtained readily from a table of values of $\frac{1}{2}(1+\alpha)$?

59. Least Squares.—If we assume a given distribution of residuals or errors to be normal, the probability of the occurrence of a given error (x_r-x), where x_r denotes the numerical value of the particular observation and x denotes the theoretical value according to a given hypothesis, is

$$p_r = ke^{-a(x_r-x)^2},$$

where k and a are constants with which we should now be familiar. Then, if the occurrence of each error is independent of the occurrence of any other error, and all the errors are to be regarded as of equal weight, the probability P of the joint occurrence of all the errors is the product of all the corresponding values of p_r or,

$$P = Ke^{-a\{(x_1-x)^2+(x_2-x)^2+\text{etc.}\}},$$

where K denotes the product of all the values of k and is, of course, constant. If the errors were not to be regarded as of equal weight the value of a would evidently vary accordingly.

The probability P is evidently greatest when the sum of the squares of the errors is the least; therefore, that distribution of errors is most likely which makes P have the greatest value or the sum of whose squares is a minimum. This principle, known as the principle of *least squares*, is very valuable as a

basis for the solution of many important problems. As a complete explanation of the possible application of the method of least squares would require the development of a certain amount of technique, and as this development is given in almost any one of several treatises devoted entirely to that subject, we shall restrict our attention to those applications which will prove sufficient for the purposes of this course.

The general form of application of the method of least squares which we shall consider concerns itself with the fitting of polynomials of the form $y=a+bx+cx^2+$etc., to given pairs of values. For this purpose it is well to note that the errors or residuals mentioned above will be the differences between the observed values of y and the corresponding theoretical values of y given by the polynomial, and the general problem consists in determining the values of the coefficients a, b, c, etc., which will make the sum of the squares of the differences between the corresponding values of y a minimum. The determination of these coefficients constitutes a problem in minima where, however, partial differentiation is usually necessary. A simple example should make the details of the general solution clear. Let us fit the linear expression $y=a+bx$ to the three pairs of coordinates (1, 2), (3, 9) and (5, 14). The residuals are then $a+b-2$ (the value of $a+bx$ for $x=1$, minus the observed value 2), $a+3b-9$ and $a+5b-14$ and we are to determine for what values of a and b the sum of the squares of these residuals is a minimum. It is easily verified that the derivatives of

$$(a+b-2)^2+(a+3b-9)^2+(a+5b-14)^2,$$

with respect to a and b reduce to $3a+9b-25$ and $9a+35b-99$ respectively. Setting these two expressions equal to zero and solving simultaneously, we obtain the required values of a and b; since these equations are the same as those obtained for the same problem by the method of moments, the values of a and b must be the same as found previously or $a=-\frac{2}{3}$ and $b=3$, and the final equation must be $y=3x-\frac{2}{3}$.

EXERCISES

For suggestions for solving the following exercises see similar exercises under Moments.

1. The following results were obtained from an experiment on a screw-jack.

R	10	19	29	46	51	66	78	89	101	113
E	.5	1	1.5	2	2.5	3	3.5	4	4.5	5

Determine a linear relation $R=aE+b$ between the effort E and the load R.

2. In the following data, which are known to follow a law represented approximately by $R=aE+b$, there are errors of observation.

R	9.5	12.5	15.5	18.5	21.5	24.5	27.25	30.5	33.5	36.5
E	4	5	6	7	8	9	10	11	12	13

Determine the most probable values of a and b.

3. In a tensile test of a mild steel bar the following observations were made, where W represents the load in tons and x the elongation in inches. (The bar had an initial length of 8 inches and a diameter of 0.748 inches.)

W	1	2	3	4	5	6
x	0.0014	0.0027	0.0040	0.0055	0.0068	0.0082

Determine the relation $x=aW+b$.

4. A wire under tension is found by experiment to stretch an amount L in thousandths of an inch under a tension T in pounds as follows:

T	10	15	20	25	30
L	8	12.5	15.5	20	23

Determine a relation of the form $L=kT$ (Hooke's Law).

5. A restaurant keeper finds that if he has G guests a day his total daily expenditure is E dollars and his total daily receipts are R dollars The following data are averages obtained from the books.

G	210	270	320	360
E	16.7	19.4	21.6	23.4
R	15.8	21.2	26.4	29.8

Determine the relations $R=mG$ and $E=aG+b$. What are the interpretations of m, a and b? Below what value of G does the business cease to be profitable?

6. If a body slides down an inclined plane, the distance S in feet that it moves in t seconds after it starts is represented by the equation $S=kt^2$. Determine the best value of k consistent with the following data:

S	2.6	10.1	23	40.8	65.7	
t	1	2	3	4	5	*Ans.* $k=2.56$.

7. An alloy of tin and lead containing x per cent of lead melts at the temperature y (F.) given by the values:

x	25	50	75
y	482	370	356

Determine the relation $y=a+bx+cx^2$.

8. The weight of ten liters of water was found at different temperatures T (C.) and the loss in weight (in g.) W as the temperature differed from 4° as follows:

W	1.3	0.3	0	0.3	1.2	2.7	4.8	7.3
T	0	2	4	6	8	10	12	14
W	10.3	13.8	17.7	22.0	26.8	31.9	37.1	43.3
T	16	18	20	22	24	26	28	30

Determine the relation $W=a+bT+cT^2$.

9. It is claimed that if the brake mechanism is satisfactory and road conditions are average, any automobile should stop at distances and speeds as follows:

Speed per hour, v....	10	15	20	25	30	35	40	50
Distance in feet, d....	9.2	20.8	37	58	83.3	104	148	231

Determine the relation $d=av^2+bv+c$.

10. Observations upon the corresponding temperature T (F.) and pressure P of steam in units of 10 were made as follows: (Atmos. pressure = 14.7 lbs.)

T	240.0	259.2	274.3	286.9	297.8
P	1	2	3	4	5
T	307.4	316.0	323.9	331.1	337.8
P	6	7	8	9	10

Compute the differences and estimate the appropriate degree of the polynomial to be fitted. Determine the polynomial.

11. The following data give the velocity of water (feet per sec.) and head in units of 10 feet.

V	25.4	35.9	43.9	50.7	56.7	62.1	67.1	71.8	76.1	80.3
H	1	2	3	4	5	6	7	8	9	10

Determine the relation $V=aH^2+bH+c$ between velocity V and head H.

12. A strong rubber band stretched under a pull of x kg., shows an elongation of y cm., as given by the observations:

x	.5	1	1.5	2	2.5	3	3.5	4	4.5	5
y	.1	.3	.6	.9	1.3	1.7	2.2	2.7	3.3	3.9

Determine the relation $y=kx^n$. *Ans.* $y=.3x^{1.6}$.

13. The intercollegiate track records for foot-races are as follows, where d is the distance run and t the record time.

d	100 yds.	220 yds.	440 yds.	880 yds.	1 mile	2 miles
t	$0:09\frac{4}{5}$	$0:21\frac{1}{5}$	0:48	$1:54\frac{4}{5}$	$4:15\frac{2}{5}$	$9:24\frac{2}{5}$

Determine a relation of the form $t=kd^n$. What should be the record time for a race of 1320 yds.?

14. The corresponding ages in years and diameters in inches of a tree with an initial height of $1\frac{1}{2}$ feet were found to be as follows:

Age, y	19	58	114	140	181	229
Diameter, x.....	3	7	13.2	17.9	24.5	33

Determine the relation $y=kx^n$.

15. Vapor pressures, in mm. of mercury, of methyl alcohol at various temperatures were found by experiment to be:

t	0	5	10	15	20	25	30	35
P	30	40	54	71	94	123	159	204

Determine an appropriate relation.

16. The temperature of a heated body cooling in the air was taken each minute, the results being tabulated as follows:

t	0	1	2	3	4	5
T	84.9	79.9	75.0	70.7	67.2	64.3
t	6	7	8	9	10	
T	61.9	59.9	57.6	55.6	53.4	

The temperature of the air was 20 degrees. Determine an appropriate relation between the temperature T and the time t.

60. Probable Error in a Single Observation. A Rough Method of Computation.—We have shown why the standard deviation may be used to measure the consistency of a set of observations. A slightly better measure of such consistency is given, however, by what is called the *probable error* in a single observation. If a distribution is normal the probable error may be defined as the deviation from the mean which, taken with both the positive and negative signs, constitutes the limits of one-half of the total frequency. If the probable error so defined is, say $\pm k$, and if the distribution is representative, the probability of another deviation selected at random falling between $-k$ and $+k$ equals the probability of its falling without that interval. A rough method of dealing with a distribution is to compute the mean, say α, of the positive deviations, and the mean, say β, of the negative deviations, and to employ the arithmetic average of the absolute values of these means as the probable error, or the following relation:

Approximate probable error in a single observation

$$= \pm \frac{\alpha - \beta}{2} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (47)$$

Such a rough application of the definition of the probable error in a single observation can be expected to give only roughly approximate results and is introduced here mainly to help clarify the conception of probable error; little refinement in the calculation should then be employed. As an example, let

us consider the distribution of litters of mice considered earlier in this course.

Number in Litter	Frequencies, y	Deviations, x	xy
1	7	−4	− 28
2	11	−3	− 33
3	16	−2	− 32
4	17	−1	− 17
5	26	0	−110
6	31	1	31
7	11	2	22
8	1	3	3
9	1	4	4
	121		60

For purposes of approximation it will be well to regard one-half of the frequency corresponding to the deviation of zero as belonging to the positive deviations and one-half to the negative deviations. Then

$$\alpha = \frac{60}{57} = 1.0,$$

and

$$\beta = \frac{-110}{64} = -1.7.$$

Hence, the probable error in a single observation is approximately $\pm\frac{1}{2}(1.0+1.7)$ or ± 1.3. If this distribution were representative, the probability of the deviation corresponding to another litter selected at random falling between -1.3 and 1.3 would equal approximately the probability of its falling without that interval. It can scarcely be emphasized too much that the principal means of obtaining a *representative* distribution is to make a large number of observations. If, however, a distribution includes a sufficiently large number of cases to be representative, the method of computing the probable error given in the next section should be employed.

EXERCISES

Find a rough approximation of the probable error in a single observation for the distribution of:

1. Cephalic indices of Bavarian skulls.

2. The results of shooting 1000 times at a vertical line.

3. Enter the table of areas of the normal curve given previously and estimate the value of x/σ which corresponds to the probable error.

61. Probable Error in a Single Observation. The Standard Method of Computation.—Most distributions call for a more refined method of computing the probable error than the one considered in the preceding section; and even in distributions which involve too few observations to be representative, the more refined method which we shall now consider will give just as good results if the ordinary rules for numerical computation are observed. If we enter the table of areas of the normal curve backward with $\frac{1}{2}(1+\alpha)=0.75$ we find that $x/\sigma=0.6745\ldots$. It follows then that one-half of the total frequency of *any* normal distribution lies between $x=-0.6745\sigma$ and $x=+0.6745\sigma$. We write then that the

$$\text{probable error in a single observation} = 0.6745\sigma \ldots . \quad (48)$$

Thus, in the case of the distribution of litters of mice, where $\sigma=1.75$, the probable error in a single observation becomes $\pm 0.6745(1.75)$ or about ± 1.18.

The probable error is one of the most valuable tools of the statistician, or even of the scientific investigator in general and, as will be indicated in a later article, can be adapted and applied to many forms of measurement. It is all very well to know, say, the value of the arithmetic average of a large number of measurements, but such a value would be worth much more if it were attended by the value of the probable

error, giving information in regard to the consistency of all of the observations. Thus, since the average size of the litters of mice was found to be 4.59 the result would be of greater value if written $4.59(\pm 1.18)$.

The value of the standard deviation is frequently used instead of the probable error for much the same purpose and is usually referred to in that case as the *standard error*. It has been found by experience that the occurrence of a deviation of more than " three " times the probable error is either very unlikely or due to peculiar influences not covered by the investigation. Thus, one would be justified in concluding that a size of a litter of mice of $4.59+3(1.75)$ or of 10 and over would be very unlikely if the distribution cited above is representative. Deviations are found occasionally, however, which seem to be fairly normal and yet which exceed slightly three times the probable error; on the other hand, such deviations practically never exceed three times the mean error and hence there are a few authorities who prefer to employ the latter criterion in testing a given deviation. We shall take the liberty of referring here to " three times the probable error " as the *maximum probable error* and " three times the standard error " as the *maximum error*, for purposes of distinction, although where no distinction is necessary we shall follow the general custom and use the maximum probable error.

EXERCISES

Compute the probable error in a single observation for the following distributions and note whether any deviations exceed the maximum probable error:

1. Of cephalic indices of Bavarian skulls.

2. Of the results of shooting 1000 times at a vertical line.

3. Of the statures of Cairo-born Egyptians (given in a preceding exercise).

4. Of the statures of Smith College girls (given in a preceding exercise).

62. Field Artillery.—While little interest will probably be felt in the subject of field artillery itself, it offers opportunities for good but simple applications of the idea of probable error which should add familiarity and confidence in the use of that idea.

The following abridged table of probable errors was taken from a much more complete table based upon the firing of over 5000 rounds with the 3-inch gun.

PROBABLE ERRORS (All in yards)

Range	Range	Vertical	Deflection
1500	39	1.8	3.1
2000	34	2.4	4.4
2500	31	3.2	5.6
3000	29	3.9	7.0
3500	28	4.9	8.6
4000	27	5.8	10.4
4500	26	6.8	12.1
5000	25	7.8	14.0

If we speak of the 50 per cent zone as that determined by the probable error, the width of a zone of any per cent can easily be obtained from the table of areas of the normal curve, given previously. We shall refer to the ratio of the width of any zone to the width of the 50 per cent zone as a *probability factor*; a table of these factors can easily be set up, and will prove useful in connection with the values of the probable errors given in the table above. An abridged table follows:

Per Cent	Probability, Factor
10	0.18
20	0.38
30	0.57
40	0.78
50	1.00
60	1.25
70	1.54
80	1.90
90	2.44

The values of probable errors for any range not given in the upper table and the values of probability factors for any per cent not given in the lower table can easily be determined by ordinary interpolation. It scarcely needs to be added that complete tables would be available and that little interpolation would be necessary in actual artillery fire.

As an example illustrating the use of the tables given above, suppose that it is desired to know the location of the center of impact (the center of gravity in range of the points of fall or impact) with respect to the target, when out of 12 shots one shot is observed to be "short" and 11 are "over" at range 2200. Since $8\frac{1}{3}$ per cent of the shots are short we are interested in the width of the $83\frac{1}{3}$ $(=100-2\times 8\frac{1}{3})$ per cent zone. Entering the table of probability factors with $83\frac{1}{3}$ per cent, we interpolate 2.05 for the probability factor. The width of the $83\frac{1}{3}$ per cent zone is then 2.05 times the width of the 50 per cent zone at 2200 yards (which according to the table of probable errors is 2×32 or 64) or 131 yards. The center of impact is then probably $65\frac{1}{2}$ (one half of 131) yards beyond the target. A battery commander would be justified by such a large result (or any result appreciably greater than 25 yards) in shortening his range.

Another problem of considerable importance is illustrated as follows: Assuming that a piece is correctly laid, what is the probability of obtaining a hit upon a target 2 yards high and 4 yards wide at a range of 3000 yards? At 3000 yards the width of the vertical 50 per cent zone is $7.8(=2\times 3.9)$ yards, and therefore the value of the probability factor is $2\div 7.8=0.26$ which (according to the table of probability factors) corresponds to 14 per cent; that is, 14 per cent of the shots will take effect in the long run between the horizontal lines drawn along the upper and lower edges of the target.

Likewise, the width of the 50 per cent zone in deflection is 14.0 yards and the value of the probability factor is $4\div 14.0=0.29$, which corresponds to 16 per cent; that is, 16 per cent of the shots will take effect in the long run between two vertical

lines drawn through the lateral edges of the target. Therefore, $0.14 \times 0.16 = 0.0224$ is the probability of a shot taking effect in the rectangle formed by the two pairs of lines, or upon the target. That is, one shot in about 45 will in the long run be a hit.

EXERCISES

1. In shooting 10 times at a tower 3000 yards distant, 9 shots were observed to go to the right and one to the left. How much of a deviation from the target was indicated?

2. Show that if A shots are observed to go to one side (i.e., "short," to the left, or below, etc.) of a target and B shots to the other side, the change to be made in the lay of the piece is the

$$\text{corresponding prob. error} \times \text{prob. factor of } \frac{|A-B|}{A+B},$$

for the given range.

3. In shooting 15 times at a house 2500 yards distant, 13 shots were observed to go "over" and two "under." How much of a deviation from the target vertically was indicated?

4. In shooting at a certain small object 3000 yards distant, 9 shots were observed to go to the right and above the target, and 1 to the left and below the target. What two changes to be made in the lay of the piece were indicated?

5–9. Assuming that a 3-inch gun is properly laid, find the number of shots which would be necessary in the long run:

(**5**) To drop a shell in a trench 3 yards wide and at a distance of 3000 yards, if the trench runs perpendicularly to the line of fire.

(**6**) To hit a tower 4 yards wide and 2800 yards distant.

(**7**) To hit a house 10 yards high and 12 yards wide at 3000 yards.

(**8**) To drop a shell in a gun emplacement or hole 12 yards long and 10 yards wide (measured in the direction of the line of fire) at 2500 yards.

(**9**) To sweep a canal, 15 yards wide, which runs quite a distance in the line of fire at 4000 yards.

10. Show that the probability of hitting between two parallel lines D yards apart at a given range is given by the per cent correspond-

ing to the probability factor "$\frac{D}{2} \div$ the corresponding probable error" for that range. The expression "corresponding probable error" is used because the parallel lines may be taken in any one of at least three directions.

63. Probable Errors of Other Quantities or Expressions.— We have been careful to refer to the probable error considered in the preceding sections as the probable error *in a single observation.* There are formulas also for determining the probable error of values of various kinds of expressions, such as the mean, the standard deviation, etc., all of which depend fundamentally upon the formula for the probable error in a single observation. A few of these formulas will now be given but without derivations. Other formulas will be introduced later.

P. E. in the mean

$$= \pm \frac{0.6745\sigma}{\sqrt{n}}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (49)$$

P. E. in the standard deviation

$$= \pm \frac{0.6745\sigma}{\sqrt{2n}}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (50)$$

P. E. in the coefficient of variation v

$$= \pm 0.6745 \frac{v}{\sqrt{2n}} \left[1 + 2\left(\frac{v}{100}\right)^2\right]^{\frac{1}{2}}. \quad . \quad . \quad . \quad . \quad (51)$$

P. E. in the observed probability p

$$= \pm 0.6745 \sqrt{\frac{pq}{n}}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (52)$$

P. E.[4] in a single observation corresponding to the sum or difference of several independent variables

$$= \pm 0.6745 \sqrt{\sigma_x^2 + \sigma_y^2 + \text{etc.}} \quad . \quad . \quad . \quad . \quad . \quad (53)$$

[4] See Jones' "First Course in Statistics," p. 158, for a derivation of this formula.

where, in each case, n denotes the number of observations.

As an example showing the use of formula (49), the probable error in the mean (4.59) of the distribution of litters of mice is $\pm\frac{0.6745(1.75)}{\sqrt{121}}=\pm 0.11$, which shows that it is an even chance whether the true value of the mean lies within the interval from $4.48(=4.59-0.11)$ to $4.70(=4.59+0.11)$ or without it.

As an example showing the use of formula (52) let us consider the probability of dying within one year at age 24, or 0.008, as given by the American Experience table of mortality used in connection with life insurance. If the value of this probability were based upon the observation of 1000 lives at that age the probable error in the value of the probability 0.008 would, according to formula (52), be

$$\pm 0.6745\sqrt{\frac{(0.008)(0.992)}{1000}},$$

or about ± 0.002, and it would be about an even chance whether the true value of the probability would be between 0.006 and 0.010 or without that interval. If, however, the probability were computed from observations covering 100,000 lives at the given age it is easily verified that the probable error would be only ± 0.0002.

The formulas for the probable error in a single observation and the probable error in an observed probability represent the same important guiding principle which may be expressed in the following form which includes a convention mentioned previously: If, in a random sample of n variates the proportion of successes is p, then the *proportion* of successes in the universe from which the sample is selected will not be likely to fall outside the limits

$$p\pm 3(0.6745)\sqrt{\frac{pq}{n}},$$

and if that universe contains altogether N variates the *number* of successes will not be likely to fall outside the limits

$$Np \pm 3(0.6745)N\sqrt{\frac{pq}{n}}.$$

In particular, the number of successes in another random sample of n variates would not be likely to fall outside the limits

$$np \pm 3(0.6745)\sqrt{npq}.$$

We shall establish the latter relation later in a slightly different connection.

Attention is called to a convention which is frequently (but not universally) employed in expressing the values of probable errors. The probable error ± 1.18, expressed $4.59(\pm 1.18)$, is to be interpreted as the probable error of a single observation and is so written only with the value of the mean (such as 4.59) for obvious reasons. A probable error without parentheses is to be interpreted as the probable error of the expression which it follows. Thus, if the probable error of the mean 4.59 is ± 0.11 it may be written more compactly 4.59 ± 0.11.

The author wishes now to refer again to formula (49) and to explain that the observations cited in Art. 37 are in fact averages of each of 36 groups of 25 digits (i.e., 0, 1, 2, . . . 9) selected at random. The complete distribution of 900 digits is as follows:

Digit	Frequency	Digit	Frequency
0	95	5	80
1	96	6	82
2	93	7	72
3	105	8	90
4	91	9	96

It is easily verified that for this distribution $M = 4.38$ and $\sigma = 2.91$. Hence, according to formula (49) we should not expect any one of the 36 means cited above to vary from 4.38 by more than $3(0.6745)2.91 \div \sqrt{25}$ or 1.18. We should not

then expect any one of the means to lie outside of the interval from 3.20 to 5.56. It is evident that only one lies outside this interval. It is easily verified that no one of the means lies outside of the interval determined by the maximum error.

EXERCISES

1. The distribution of workers in two districts of England in a certain year was as follows:

	Workers Over 35 Years Old	Workers at All Ages
District A.....	11,718	35,316
District B.....	4,029	21,822

Determine whether the difference in the proportion of workers over 35 years of age for the two districts is more than one would expect in random sampling. Suggestion: Compute p the ratio of all workers over 35 to all workers and the values of the standard deviations of this ratio for each district and then employ formula (53).

2. Solve the preceding exercise in the same general way but use different values of p (i.e., 0.332 and 0.185) in computing the two values of the standard deviation. What is the difference between the two assumptions used in the two methods of computation?

3. Investigations have been made to find whether there is any significant difference in the size of eggs of cuckoos laid in general and laid in nests of certain species of foster parents. The results are as follows:

Eggs laid	Number	Mean Length (m.m.)	σ
generally	1572	22.3	0.96
in nests of			
Garden Warbler..	91	21.9	0.79
White Wagtail...	115	22.4	0.76
Hedge Sparrow...	58	22.6	0.86

Determine whether there is any significant difference in each of the three situations. Compute the mean errors of the means and employ formula (53).

4. The following results were obtained in an investigation into any significant differences in the males of a certain species of crab when found in deep water and when found in shallow water.

	Mean Carapace Length (mm.)	σ	v
Deep water	8.59 ± 0.05	1.67 ± 0.04	19.45 ± 0.44
Shallow water	8.41 ± 0.04	1.49 ± 0.03	17.75 ± 0.37

Determine any significant difference.

Ans. Means 0.18 ± 0.07 Possibly.
σ 0.18 ± 0.05 Probably.
v 1.70 ± 0.58 Possibly.

5. The standard deviation of the head-lengths of 3000 criminals was found to be 6.04593 ± 0.05265 (mm.), and of 1306 of these criminals selected at random 6.00247 ± 0.07922 (mm.). Determine whether the difference is significant. If the difference were not significant would this fact prove the whole group to be homogeneous? If no significant differences in the means and all other expressions were found, what could be concluded?

6. The standard deviations of the lengths and breadths of 139 skulls of the Naqada race, excavated in Upper Egypt and believed to be some 8000 years old, were found to be 5.722 and 4.612 mm., respectively, and the same for 1000 Cambridge undergraduates were 6.161 and 5.055 respectively. Are the differences significant?

7. The heights in inches of groups of fathers, mothers, sons and daughters were found to be as follows:

	Fathers	Mothers	Sons	Daughters
Mean	67.68 ± 0.06	62.48 ± 0.05	68.65 ± 0.05	63.87 ± 0.05
σ	2.70 ± 0.04	2.39 ± 0.04	2.71 ± 0.04	2.61 ± 0.03
v	3.99 ± 0.06	3.83 ± 0.06	3.95 ± 0.06	4.09 ± 0.05

(*a*) Determine the number of fathers, mothers, sons and daughters.

(*b*) Check the values of the probable errors.

8. A random sample of 90 was selected from 514 candidates in a certain examination where the grades ranged from 3 to 64. The grades by groups and the percentages of the whole group and of the sample making these grades were found to be as follows:

Grades	Percentages	
	Total	Sample
...14	8	8±1.9
15–24	19	17±2.6
25–29	16	18±2.7
30–34	18	13±2.4
35–39	15	17±2.6
40–49	19	18±2.7
50...	7	10±2.1

Check the probable errors in the last column. Are there any evidences of special lack of homogeneity in the whole group?

9. A sample of 60 towns was selected at random from 241 towns in England and Wales, and the number of infectious diseases per thousand of population for the sample and for the whole group of towns was tabulated to give the following results:

Rate per Thousand	Frequencies	
	All Towns	Sample
1– 4	85	92±10
5– 8	86	96±10
9–12	42	28±7
13...	28	24±6

Check the probable errors. Point out any evidences of special lack of homogeneity in the whole group. What is the practical importance of examples like this and the one given in the preceding exercise?

10. An attempt was made to predict the annual output per head (in pounds sterling) in 142 different types of employment for 1907 in the United Kingdom, from an analysis of the output of 50 different occupations selected at random to give the following results:

Output per Head	Number of Occupations		
	Sample	Predicted for Total	Actual Number
.....59	4	11±3.6	12
60– 79	16	45±6.2	42
80– 99	6	17±4.3	25
100–119	10	28±5.3	20
120–189	8	23±4.9	27
190....	6	17±4.3	16

Check the probable errors. Note any special failures in prediction.

NOTE.—The method followed in obtaining the random samples cited in the above examples is simple but irksome and should perhaps be explained. The total number of observations was ranked and numbered in each case, and then a sample of these numbers was selected by forming numbers out of the digits found, say, in the seventh decimal place of successive groups of logarithms. Thus, to obtain a sample of 50 out of 142, the seventh-place digits of each successive group of three logarithms would be arranged in the order of their occurrence and every number so obtained which exceeded 142 was ignored. The first 50 numbers retained would serve to identify the sample.

CHAPTER IX

THE BINOMIAL $(p+q)^n$. STATISTICAL SERIES

64. Asymmetrical Curves.—As may be readily inferred, there are many curves which are like the normal curve in that they have at most one mode, but which differ from the normal curve in that they are not symmetrical. Such curves are often referred to as *skew* curves. Most of these types of curves are represented by the various graphs obtained by plotting the terms of the expansion of the binomial $(p+q)^n$, where p represents the probability of the occurrence of a certain event, q the probability of its failure and n the number of trials. Lest one should think that the graphs of the terms of the expansion would be rectangular histograms, it is well to say that it is the frequency curves which correspond to these histograms to which we refer as representative types of frequency curves. When, as a special case, $p=q=\frac{1}{2}$, the corresponding graph is symmetrical—that is, skewness is absent or zero—and is essentially a form of the normal curve.

Let us consider again a few concrete illustrations of such an expansion and acquire greater familiarity with the significance of the terms of the expansion. If n coins were tossed up, the first term of the following expansion

$$(\tfrac{1}{2}+\tfrac{1}{2})^n=(\tfrac{1}{2})^n+n(\tfrac{1}{2})^{n-1}(\tfrac{1}{2})+\frac{n(n-1)}{2}(\tfrac{1}{2})^{n-2}(\tfrac{1}{2})^2+\text{ etc.},$$

is the probability that all coins will be "heads." Similiarly, the second term is the probability that all but one will be "heads"; since $(\frac{1}{2})^{n-1}(\frac{1}{2})$ is the probability that one particular coin will be "tails" and the rest "heads," and this one

coin can be chosen in n ways. Likewise, the third term is the probability that all but two will be "heads," and so on for the rest of the terms. The probability of obtaining exactly r "tails" with n coins is then ${}_nC_r(\frac{1}{2})^{n-r}(\frac{1}{2})^r$, where $(\frac{1}{2})^{n-r}(\frac{1}{2})^r$ is the probability that one particular set of r coins of the total n coins will be "tails" and the rest "heads," and ${}_nC_r$ is the number of ways these r coins can be selected from n coins; that is, ${}_nC_r$ is the number of combinations of n things taken r at a time.

Likewise, the probabilities of throwing no ace, of throwing exactly one ace, exactly two aces, etc., in throwing n dice are given by the successive terms of the expansion

$$(\tfrac{5}{6}+\tfrac{1}{6})^n=(\tfrac{5}{6})^n+n(\tfrac{5}{6})^{n-1}(\tfrac{1}{6})+\text{ etc.}$$

It should be emphasized that the terms of these expansions represent probabilities and that their sum is unity. If the terms of such expansions were multiplied all the way through by a suitable number, the various terms would obviously represent probable frequencies. Thus, the terms of the expansion

$$64(\tfrac{1}{2}+\tfrac{1}{2})^4=4+16+24+16+4,$$

represent the number of times the various possibilities "four heads," "three heads and one tail," etc., would be most likely to occur *in the long run* in tossing 4 coins 64 times. No one would, of course, expect to obtain these frequencies in a single experience, although he might come very close to it; the results of an actual experience were $5+15+24+15+5$.

EXERCISES

Compute the theoretical frequencies corresponding to the following distributions:

1. Three dice were thrown 648 times and the number of times a "5 or 6" appeared was tabulated as follows:

Number	Frequency
0	179
1	298
2	141
3	30

How many times did a " 5 or 6 " appear? How many times was it possible for a " 5 or 6 " to appear?

2. Balls were drawn, one at a time, from a bag containing an equal number of black and white balls, each ball being returned before the next drawing. The number of black balls drawn was then tabulated for each consecutive seven drawings as follows:

Number	Frequency	Number	Frequency
0	9	4	148
1	34	5	95
2	104	6	40
3	151	7	4

3–5. Balls were drawn, one at a time, from a bag containing an equal number of black and white balls. The number of black balls was then tabulated for each consecutive (3) five drawings, (4) six drawings and (5) seven drawings, to give the following distributions:

Number of Black Balls	(3)	(4)	(5)
0	30	17	9
1	125	65	34
2	277	166	104
3	224	192	151
4	136	166	148
5	27	69	95
6	...	8	40
7	...	...	4

Determine how many black balls were drawn and how many it was possible to draw in each set.

6. Three dice were thrown 196 times and the sum of their upper faces was tabulated to give the following distribution:

Number Thrown	Frequency	
4	1	*Ans.* 1
5	4	3
6	11	5
7	10	9
8	24	14
9	22	19
10	22	23
11	32	24
12	17	24
13	23	23
14	9	etc.
15	7	
16	7	
17	4	
18	3	

7. A coin was tossed 2048 times and every time a " head " appeared a new record was made to show in which toss " head " appeared for the first time. " Head " appeared for the first time in the

1st toss	1061 times
2nd "	494 "
3rd "	232 "
4th "	137 "
5th "	56 "
6th "	29 "
7th "	25 "
8th "	8 "
9th "	6 "

8. Twelve dice were thrown 4096 times and the number of times a " 6 " appeared was tabulated for each throw to give the distribution:

Number	Frequency	Number	Frequency
0	447	5	115
1	1145	6	24
2	1181	7	7
3	796	8	1
4	380		

9. The appearence of a " 4, 5 or 6 " was tabulated for 4096 throws of 12 dice to give the distribution:

Number	Frequency	Number	Frequency
0	0	7	847
1	7	8	536
2	60	9	257
3	198	10	71
4	430	11	11
5	731	12	0
6	948		

10. The following distribution gives the number of trumps held by the first hand in 25,000 deals at whist:

Number of Trumps	Frequency	Number of Trumps	Frequency
0	215	5	2950
1	1724	6	852
2	5262	7	166
3	7440	8	20
4	6371		

Show that the number of deals which would be necessary, in the long run, to yield a hand consisting entirely of trumps would be almost seventeen million.

11. Fourteen coins were tossed 150 times and the number of heads and corresponding frequencies was tabulated to give:

Number	Frequency	Number	Frequency
3	2	8	25
4	15	9	15
5	17	10	6
6	27	11	6
7	36	12	0
		13	1

12. Eight dice were thrown 6561 times, and the number of times a " 5 or 6 " appeared was tabulated to give the distribution:

Number	Frequency	Number	Frequency
0	256	5	448
1	1024	6	112
2	1792	7	16
3	1792	8	1
4	1120		

13. The number of occurrences of any one of 0, 1, 2, 3 or 4 in the seventh decimal place was tabulated for each of 300 groups of 50 logarithms to give the distribution:

Number	Frequency	Number	Frequency
14	1	25	42
15	0	26	36
16	3	27	30
17	2	28	28
18	3	29	15
19	7	30	16
20	9	31	5
21	18	32	2
22	26	33	2
23	21	34	1
24	32	35	1

14. Eleven dice were thrown 26,306 times and the number of times a " 5 or 6 " appeared was tabulated to give the distribution:

Number	Frequency	Number	Frequency
0	185	6	3067
1	1149	7	1331
2	3265	8	403
3	5475	9	105
4	6114	10	14
5	5194	11	4

15. Twelve dice were thrown 10,596 times and the number of times a " 4, 5 or 6 " appeared was tabulated to give the distribution:

Number	Frequency	Number	Frequency
0	1	7	2198
1	21	8	1380
2	163	9	648
3	500	10	188
4	1141	11	32
5	1962	12	3
6	2359		

65. The Probable Error in a Single Observation and the Expansion of $(p+q)^n$.—The terms of the expansion $(p+q)^n$ form the basis of interesting investigations. If we expand

the binomial in the order indicated by $(q+p)^n$ and multiply the first term q^n by 0, the second term $nq^{n-1}p$ by 1, the third term by 2, and so on, the sum of all such products is the first moment *about* q^n; and since $(q+p)^n=1$, the moment is a unit moment. Hence,

$$\nu_1' = nq^{n-1}p+n(n-1)q^{n-2}p^2+\frac{n(n-1)(n-2)}{2}q^{n-3}p^3+\ldots+np^n$$

$$=np\{q^{n-1}+(n-1)q^{n-2}p+\ldots+(n-1)qp^{n-2}+p^{n-1}\}$$

$$=np(q+p)^{n-1}$$

$$=np.$$

It is left for the student to show that the second unit moment about q^n is

$$\nu_2' = np+n(n-1)p^2.$$

But the second unit moment about the mean or

$$\nu_2=\nu_2'-(\nu_1')^2$$

$$=np+n(n-1)p^2-n^2p^2=np-np^2=np(1-p)$$

$$=npq.$$

Therefore, the standard deviation or

$$\sigma=\sqrt{npq} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (54)$$

and the probable error in a single observation

$$=0.6745\sqrt{npq}. \quad . \quad . \quad . \quad . \quad . \quad . \quad (55)$$

Let us consider the significance of these results. Since $\nu_1'(=np)$ is the distance from the term q^n to the mean, the np-th term measured from q^n is at the mean and if we regard the mean or arithmetic average as the most probable value, the np-th term measured from q^n represents the term of greatest probability.[1] Now the np-th term measured from

[1] As a matter of fact, it can be shown that the value of this term is the greatest. See Fisher's "Mathematical Theory of Probabilities," p. 100.

q^n represents the probability that the event under consideration will happen exactly np times in n trials; we conclude, then, that np is the most probable number of times that the event will occur in n trials. This is, of course, what we should expect; for example, if we toss up 400 coins the most probable number of "heads" is $np=400\cdot\frac{1}{2}=200$. It is fundamentally important to think of np as the most probable number and not as the corresponding probability (that is, as an abscissa and not as an ordinate).

The probable error, say $k(=0.6745\sqrt{npq})$, is the divergence (an abscissa) on either side of the most probable number (the mean) such that the frequency or area between $x=-k$ and $x=+k$ is approximately (*exactly* if the distribution were truly normal) half of the total frequency or area; and the probability of an observation chosen at random falling within the interval is approximately the same as that of its falling without the interval. In the illustration given above (of tossing 400 coins) the probable error is $0.6745\sqrt{400\tfrac{1}{2}\cdot\tfrac{1}{2}}$ or about 7; therefore, a deviation of more than three times 7 from 200 "heads" is not to be expected if attendant circumstances are normal. A deviation of 50 from 200 "heads" would practically establish the existence of abnormal conditions, such as "influenced tossing," a "loaded coin," etc. If the value of p used in such an investigation were based solely upon experience (such as a death rate) a deviation of more than three times the probable error would lead naturally to the conclusion that the value of p used in the investigation was not representative. As a final illustration, suppose that it had been found by experience that the ratio of male children born to female children born was 1050/1000; in other words, the probability of a child being male is 1050/2050. If 5135 out of 10,000 children proved to be males, in a certain community, what conclusions could be drawn from the deviation from the expected number? It is easily verified that the expected number is 5122 and that the probable error is ± 33.7. The deviation of 13 is then well within the value of the prob-

able error and is to be expected. A divergence of, say 300 would, however, show that the given ratio did not fit the given situation.

EXERCISES

1. Determine whether the following actually observed results were to be expected:

(*a*) 2048 " heads " in 4040 throws of a coin;

(*b*) 4096 " heads " in 8132 throws of a coin;

(*c*) 10,353 " heads " in 20,480 throws of a coin;

(*d*) 2030 black balls in 4096 drawings of a single ball from an urn containing black balls and white balls in equal proportions;

(*e*) same as (*d*) but 8120 black balls in 16,384 drawings;

(*f*) 81,236 trumps in 325,000 deals of a single card;

(*g*) 39,756 appearances of a " 4, 5 or a 6 " in 78,000 throws of a single die;

(*h*) 106,602 appearance of a " 5 or 6 " in 289,366 throws of a single die;

(*i*) 7513 appearances of a " 0, 1, 2, 3 or 4 " in a certain decimal place in 15,000 logarithms;

(*j*) 64,901 appearances of a " 4, 5 or 6 " in 127,052 throws of a single die;

(*k*) 470 appearances of a " 4 or 6 " in 1944 throws of a single die;

(*l*) 7440 holdings of three trumps in 25,000 deals at whist;

(*m*) in 32 times out of 196 throws of three dice the sum of the upper faces were 11;

(*n*) a " 5 or 6 " appeared twice 141 times in 648 throws of three dice;

(*o*) a " 6 " failed to appear 447 times in 4096 throws of 12 dice.

2. The number of births by sexes in Denmark for the designated quinquennial periods were as follows:

	Female	Male
1860–4	130,089	138,289
1865–9	135,324	142,828
1870–4	139,733	148,360
1875–9	154,214	162,823

Determine whether the deviations of the number of males from the number determined by the average ratio of males to females are to be expected.

3. The number of twins were found for each of five genealogical records to be as follows:

	Number of Births	Twins
1	4184	11
2	4116	32
3	4147	11
4	3421	25
5	4744	8

Determine the empirical probability of a birth resulting in twins, and whether any number of twins given above is unexpected.

4. The number of accidents resulting in permanent disability, and the number of fatal accidents, for various countries for specified periods were as follows:

		Permanent Disability	Fatal
Austria	(1897–1906).....	82,446	8,349
Belgium	(1905–1908).....	8,204	1,838
Denmark	(1899–1906).....	4,192	389
France	(1899–1908).....	140,877	18,708
Germany	(1899–1908).....	313,219	59,893
Italy	(1898–1902).....	9,701	2,224
Norway	(1895–1905).....	4,496	832
Russia	(1904–1906).....	34,981	2,345

Compute the average ratio of the number of fatal acidents to the number of accidents resulting in permanent disability, and find whether the relative number of fatal accidents in the case of either country listed above is unexpected.

5. Assuming the presence of no abnormal circumstances, determine the most probable number of occurrences of the events designated in the following examples and the maximum deviation therefrom to be expected (i.e., the maximum probable error):

(a) the number of heads in tossing a coin 10,000 times;

(b) the number of aces in 10,000 throws of a single die;

(c) the number of shots taking effect in the first quadrant in shooting 10,000 times at a target consisting of the origin of a pair of rectangular axes;

(d) same as (c) except that the axes are oblique and at an angle of 60°;

(e) the number of aces in 1600 drawings of a single card from a pack of cards;

(f) the number of times a marble falls to one side of a line used as a target in dropping the marble 5000 times;

(g) number of times a red ball is drawn from a bag containing 5 balls, each of a different color, in 10,000 drawings of a single ball;

(h) same as (g) except that the 5 balls consist of 2 reds and 3 blacks;

(i) the number of deaths per annum out of 100,000 individuals of a given age, where the death rate for that age is known to be 0.00864;

(j) the number of male births out of a total of 100,000 births, assuming the ratio of male births to female births to be 1050 : 1000;

(k) the number of deaths in a general population of 1,000,000, where the death rate is known to be 14 per thousand.

6. If the probability of the occurrence of a certain event in a single trial is p, show that the probability that the event will occur np times in n trials is

$$\frac{n!}{(np)!\,(nq)!}\,p^{np}q^{nq}.$$

7. The following formula is known as Stirling's formula and is employed to obtain approximations of factorials of high order:

$$n! = n^{n+\frac{1}{2}}e^{-n}\sqrt{2\pi}.$$

Show that the value of the probability given in the preceding exercise may be written

$$\frac{1}{\sqrt{2\pi npq}}.$$

8. It has been found that the average deviation of a distribution with a constant probability as a base (called a Bernoullian distribution) is

$$2n\frac{n!}{(np)!\,(nq)!}\,p^{np+1}q^{nq+1}.$$

Show that this expression may be written

$$\sqrt{\frac{2}{\pi}}\cdot\sqrt{npq}=\sigma\sqrt{\frac{2}{\pi}}.$$

9. Show that the third unit moment of the expansion of $(q+p)^n$ about the mean is $npq(q-p)$.

10. Show that the fourth unit moment of the expansion of $(q+p)^n$ about the mean is $npq\{q^2+(3n-4)pq+p^2\}$.

66. Bernoullian, Poisson and Lexian Series.—Heretofore we have given no consideration to the possibilities of breaking up a group of observations into several sub-groups or sets for individual investigation and comparison. Such a consideration will lead to some very important conclusions. Let us think of a large number of observations as classified into N sets of equal size and consider the three following situations:

A. Suppose that the probability of the occurrence of a certain event remains constant during each and all of the N sets. Suppose that the number of times the event occurs in the first set is r_1; in the second set r_2; and so on for all the sets. Then the series of absolute frequencies thus obtained is called a *Bernoullian series.*

B. Suppose that the probability varies from trial to trial within each set, but that the values and the variations are the same for all sets. The corresponding series of frequencies is called a *Poisson series.*

C. Suppose that the probability remains constant within each set but varies from set to set. The corresponding series of frequencies is called a *Lexian series.*

The distinctions made above may be illustrated as follows: Suppose that each of N bags contains black and white balls in the same proportion, and that n balls are drawn, one at a time, from each of the bags and the color noted, each ball being returned before a new drawing is made. If the number of black balls drawn from the first bag is r_1, the number from the second bag is r_2, and so on, the number sequence

$$r_1, r_2, r_3, \ldots r_N$$

is a Bernoullian series.

Suppose, however, that the proportion of black and white balls is changed each time a ball is drawn in each set, but that the program is exactly the same for each set; the number sequence of black balls drawn is then a Poisson series.

If the proportion of black and white balls is the same throughout each set but is altered from set to set, the number sequence of black balls drawn is a Lexian series.

An analysis and comparison of the means and the dispersions of the three types of series just considered lead to important results. These results will be developed in the following sections.

67. Dispersion of Ratios.—The observing student will soon become familiar with the fact that the results of most statistical investigations are better expressed in the form of ratios. Final results generally mean more if so expressed. It will be recalled that the values of moments mean little unless the moments are unit moments, and that the method suggested for computing unit moments is equivalent to dividing the frequencies of a distribution by the total population and working with the ratios. Let us consider, then, the means and the dispersions of the three types of series defined above where, however, we deal with ratios.

We shall let p_i denote either the ratio or probability corre-

sponding the i-th observation of each set of a Poisson series, or the ratio or probability corresponding to the i-th set of a Lexian series. We shall then denote the arithmetic average of these ratios or probabilities in the case of either type of series by p.

A. Bernoullian Series.—As the basic probability p remains the same, not only from observation to observation but also from set to set, the square of the dispersion for each set, and therefore for all sets as a whole, must be npq. Denoting the dispersion of a Bernoullian series by σ_B, we have

$$\sigma_B^2 = npq. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (56)$$

B. Poisson Series.—As the basic probability changes from observation to observation but the program is the same for all sets, the square of the dispersion is the same for each set and hence for all sets. The square of the dispersion for the i-th observation of each set is p_iq_i and, according to formula (53), the square of the dispersion of the whole set is the sum of all such expressions or Σp_iq_i from $i=1$ to $i=n$ inclusive. But

$$p_i = p+(p_i-p),$$

$$q_i = q-(p_i-p),$$

Hence

$$p_iq_i = pq-(p_i-p)(p-q)-(p_i-p)^2,$$

whence it is easily verified that

$$\Sigma p_iq_i = npq-\Sigma(p_i-p)^2$$

Denoting the dispersion of a Poisson series by σ_P we have then

$$\sigma_P^2 = \sigma_B^2-\Sigma(p_i-p)^2, \quad . \quad . \quad . \quad . \quad . \quad (57)$$

where the summation is to extend from $i=1$ to $i=n$ inclusive.

C. Lexian Series.—Since the basic probability for the i-th

set is p_i, the square of the dispersion about the mean np_i of that set, for that set, is np_iq_i. But we wish to take our origin at np, the mean of all the N sets. Hence, reversing the usual process (i.e., moving the origin away from instead of to the mean), we find that the square of the dispersion of the i-th set (about the mean of all the sets) is $np_iq_i+(np_i-np)^2$. A reasonable estimate for the square of the dispersion for all N sets is then the arithmetic average of the values of the expression just found for all the sets. Denoting the dispersion of a Lexian series by σ_L, we find that

$$\sigma_L^2=\frac{n}{N}\Sigma p_iq_i+\frac{n^2}{N}\Sigma(p_i-p)^2,$$

where the summation is to extend from $i=1$ to $i=N$ inclusive.

But it is evident, from a similar summation considered in connection with the dispersion of the Poisson series, that

$$\Sigma p_iq_i=Npq-\Sigma(p_i-p)^2.$$

Hence, we have finally

$$\sigma_L^2=\sigma_B^2+\frac{n^2-n}{N}\Sigma(p_i-p)^2. \quad . \quad . \quad . \quad . \quad (58)$$

where, to repeat, the summation is to extend from $i=1$ to $i=N$ inclusive.

Let us consider the means. The mean of a Bernoullian series is evidently

$$M_B=np.$$

Also, since

$$p=\frac{1}{n}(p_1+p_2+\ \ldots\ +p_n),$$

or

$$\frac{1}{N}(p_1+p_2+\ \ldots\ +p_N),$$

according as the series is Poisson or Lexian, the mean of a Poisson series, which is the same as the mean of each set, is

$$M_P = \frac{np_1 + np_2 + \ldots + np_n}{n}$$
$$= np,$$

and the mean of a Lexian series, which is the average of the means of the various sets, is

$$M_L = \frac{np_1 + np_2 + \ldots + np_N}{N}$$
$$= np.$$

We find, then, that the mean of a Poisson series is the same as the mean of a Bernoullian series and that the mean of a Lexian series is the same as the mean of a Bernoullian series whenever the probability p of the Bernoullian series is the average of the probabilities of the other type of series.

The important fact to be noted, however, is that the dispersion of a Bernoullian series is less than that of the corresponding Lexian series and greater than that of the corresponding Poisson series. This fact is evident from a mere inspection and comparison of the formulas for the dispersions of Poisson and Lexian series, but the fact is so important that the student should verify the statement to his complete satisfaction.

68. Numerical Examples of Poisson and Lexian Series.— We shall now consider two numerical examples of series which should serve to add familiarity with the processes indicated by the formulas derived in the preceding section.

As an example of a Poisson series, let us consider the following series or distribution, which was obtained by making 100 sets of 9 drawings of a single ball from a bag containing black and white balls. In each set the first drawing was made when there were 9 black and 1 white balls in the bag, the second drawing when the proportion was 8 to 2, the third 7 to 3, etc.; the proportion in the last drawing was 1 to 9. The number of black balls drawn in each successive 9 drawings was then tabulated for all 100 sets, where, of course, the program was the same for each of the sets.

Poisson Series.—Number (m) of black balls and the frequencies (y) of these numbers in 100 sample sets of 9 drawings.

$$n=9,\ N=100.$$

m	y	x	xy	x^2y
1	2	−4	− 8	32
2	4	−3	−12	36
3	13	−2	−26	52
4	26	−1	−26	26
5	32	0		
6	17	1	17	17
7	5	2	10	20
8	1	3	3	9
	100		−42	192

The correction to the mean is then -0.420 and the mean is at $M=5.000-0.420=4.580\pm0.089$,

where the value of the probable error is added to show how much of a deviation would naturally be expected as a result of chance.

Likewise $\sigma^2=1.920-(0.420)^2$,

and $\sigma=1.320\pm0.062$.

Referring now to the formulas of the preceding section, we have

$$p=\tfrac{1}{9}(\tfrac{9}{10}+\tfrac{8}{10}+\ \ldots\ +\tfrac{1}{10})=\tfrac{1}{2}.$$

Hence, $q=\frac{1}{2}$ and $np=9\cdot\frac{1}{2}=4.5$.

Therefore, $M_P=M_B=4.500$,

which evidently checks with 4.580 ± 0.089, found directly from the observed data.

It is easily verified that the value of $\Sigma(p_i-p)^2$ is $(\frac{9}{10}-\frac{5}{10})^2+(\frac{8}{10}-\frac{5}{10})^2+$ etc., or 0.600.

Hence, according to formula (57),

$$\sigma_P^2=2.250-0.600=1.650,$$

and $\sigma_P=1.285$,

which checks satisfactorily with 1.320 ± 0.062 found directly from the observed data. Also

$$\sigma_B = \sqrt{9 \cdot \tfrac{1}{2} \cdot \tfrac{1}{2}} = 1.500,$$

which is obviously greater than the value of σ_P.

We shall comment later at considerable length on the extreme rarity, in practice, of pure examples of any of the types of series under consideration. In order to prepare the student to expect departures from the rigid definitions of the three types of series, we shall consider as an example of a Lexian series one wherein the basic probability does remain the same throughout each set and does vary from set to set in some but not in all cases. The general plan of procedure will be exactly the same as if the basic probability were different for each set. The following distribution was obtained by making 90 sets of 10 drawings of a single ball from a bag containing black and white balls, and tabulating the number of black balls corresponding to each successive 10 drawings. The proportion of black balls to white balls throughout the first 10 sets of 10 drawings was 9 to 1; throughout the second 10 sets, 8 to 2; and so on until the last 10 sets, when the proportion was 1 to 9.

Lexian Series.—Number (m) of black balls and the frequencies (y) of these numbers in 90 sets of 10 drawings of a single ball. $n = 10$, $N = 90$.

m	y	x	xy	x^2y
0	6	−5	−30	150
1	6	−4	−24	96
2	7	−3	−21	63
3	9	−2	−18	36
4	8	−1	− 8	8
5	12	0		
6	10	1	10	10
7	14	2	28	56
8	5	3	15	45
9	9	4	36	144
10	4	5	20	100
	90		8	708

The correction to the mean is then $\frac{8}{90}=0.089$ and the mean is

$$M=5.000+0.089=5.089\pm0.132.$$

Likewise

$$\sigma^2=708/90-(0.089)^2,$$

or

$$\sigma=2.803\pm0.209.$$

Let us see how these results compare with the results to be obtained by the formulas of the preceding section.

$$p=\frac{10\cdot\frac{9}{10}+10\cdot\frac{8}{10}+\ldots+10\cdot\frac{1}{10}}{90}=\frac{1}{2}.$$

Hence,

$$q=\tfrac{1}{2} \quad \text{and} \quad np=10\cdot\tfrac{1}{2}=5.$$

Therefore,

$$M_L=M_B=5.000$$

which evidently checks with 5.089 ± 0.132 found from the observed data.

Also

$$\sigma_B=\sqrt{npq}=\tfrac{1}{2}\sqrt{10}=1.581.$$

Since

$$\Sigma(p_i-p)^2=10(\tfrac{9}{10}-\tfrac{5}{10})^2+10(\tfrac{8}{10}-\tfrac{5}{10})^2+\text{ etc.},$$
$$=6.000,$$

and

$$\frac{n^2-n}{N}=\frac{100-10}{90}=1,$$

we have, by formula (58)

$$\sigma_L^2=2.500+6.000=8.500,$$

or

$$\sigma_L=2.915,$$

which checks satisfactorily with 2.803 ± 0.209 found from the observed data. It is to be noted that σ_L is greater than σ_B.

The student is urged to set up examples of his own similar to those given above. It should be noticed in that connection that, if the records of the observations are kept properly, the data may be grouped to give either a Lexian or a Poisson series and one experiment he made to serve for two.

EXERCISES

Compute and compare the observed and the theoretical values of the mean and the dispersion of the following series.[2] Satisfactory comparison can be made only if the probable errors of the observed values are also computed.

1. One hundred sets of 100 drawings of a single ball from a bag containing an equal number of black and white balls were made, and the number of black balls drawn in each set was tabulated to give the following distribution:

Number of Black Balls	Frequencies		
34	1		
37	0		
40	5		
43	8		
46	15		
49	25		
52	19		
55	16		
58	8	$M=50.1\pm0.54$	$\sigma=5.33\pm0.38$
61	2		
64	1	$M_B=50.0$	$\sigma_B=5.00$

2. One thousand sets of 10 drawings of a single card from a pack of cards were made, and the number of black cards drawn in each set was tabulated to give the following distribution:

Number of Black Cards	Frequencies		
0	3		
1	10		
2	43		
3	116		
4	221		
5	247		
6	202		
7	115		
8	34	$M=4.93\pm0.05$	$\sigma=1.55\pm0.04$
9	9		
10	0	$M_B=5.00$	$\sigma_B=1.58$

[2] Most of these exercises were taken from Arne Fisher's "Mathematical Theory of Probabilities." His answers are given also.

3. One hundred sets of 10 drawings of a single card were made from a pack of cards and the number of black cards drawn in each set was tabulated to give the following distribution. During the first ten sets the proportion of black cards to red cards was 26 to 26; during the second ten sets, 25 to 27, etc.

Number of Black Cards	Frequencies		
1	4		
2	9		
3	19		
4	21	$M=4.38\pm0.17$	$\sigma=1.67\pm0.12$
5	23		
6	10	$M_L=4.14$	$\sigma_L=1.64$
7	12		
8	2	$M_B=4.14$	$\sigma_B=1.56$

4. One hundred sets of 27 drawings of a single card were made from a pack of cards, and the number of black cards drawn in each set was tabulated to give the following distribution. In each set a black card was replaced by a red card from another pack after each drawing, so that in each set the proportion of black cards to red cards were 26 to 26, 25 to 27, . . . 0 to 52 in the respective order of the drawings.

Number of Black Cards	Frequencies		
3	2		
4	6		
5	14		
6	14		
7	22		
8	17		
9	14	$M=7.16\pm0.21$	$\sigma=1.94\pm0.15$
10	8		
11	1	$M_P=6.75$	$\sigma_P=2.11$
12	1		
13	1	$M_B=6.75$	$\sigma_B=2.25$

5. Twenty sets of 500 single drawings of a card from a pack of cards were made and the number of diamonds drawn in each set was tabulated to give the following results:

109, 116, 117, 121, 122, 123, 124, 124, 129, 130, 130, 132, 133, 135, 135, 136, 138, 139, 142, 143.	$M=128.9\pm2.01$	$\sigma=8.96\pm1.42$
	$M_B=125.0$	$\sigma_B=9.68$

6. Twenty sets of 500 drawings of a single ball from an urn were made, and the number of black balls drawn in each set was tabulated. The proportion of black balls to white balls varied from set to set, 20 to 20, 19 to 21, . . . 1 to 39. The following distribution was obtained.

251	176	140	69	$M=136.6\pm15.9$	$\sigma=70.1\pm11.1$
246	183	127	55		
222	173	115	43	$M_L=131.3$	$\sigma_L=72.7$
216	156	96	29		
193	135	78	19	$M_B=131.3$	$\sigma_B=9.8$

7. One hundred sets of 100 drawings were made from a pack of cards, and the number of aces drawn in each set was tabulated to give the following distribution:

Number of Aces	Frequencies		
2	1		
3	8		
4	8		
5	7		
6	9		
7	21		
8	13		
9	15		
10	3		
11	9		
12	1		
13	2		
14	2		
15	0	$M=7.45\pm0.28$	$\sigma=2.79\pm0.20$
16	0		
17	1	$M_B=7.69$	$\sigma_B=2.66$

8. Five hundred sets of 20 drawings of a single ball from an urn were made, and the number of black balls drawn in each set was

tabulated to give the following distribution. In each set the proportion of black balls to white balls was 20 to 20, 19 to 21, etc., in the respective order of the drawings. Compute the necessary probable errors.

Number of Black Balls	Frequencies		
0	2		
1	9		
2	35		
3	52		
4	86		
5	109		
6	85		
7	69	$M=5.14$	$\sigma=1.93$
8	30		
9	16	$M_P=5.25$	$\sigma_P=1.86$
10	6		
11	1	$M_B=5.25$	$\sigma_B=1.97$

9. Ten cards were drawn 100 times from a pack of 52 cards, and the number of " black " cards tabulated for each drawing; and this procedure was repeated 26 times. After each set of 100 drawings a black card was replaced by a red card from another pack, so that the ratios of " black " to " reds " in the successive sets were as follows: 26 : 26, 25 : 27, 24 : 28, etc. The following total frequencies were obtained:

Number of Blacks	Frequencies		
0	406		
1	464		
2	454		
3	454		
4	348		
5	259		
6	130		
7	63	$M=2.625\pm0.038$	$\sigma=1.938\pm0.018$
8	17		
9	5	$M_B=2.596$	$\sigma_B=1.386$
	2600		

69. Practical Applications.—We are now ready for some practical applications of our knowledge of the relative values of the dispersions of the different types of series just considered. In the vast majority of problems occurring in practice, the values of the basic probabilities will be unknown and must be determined empirically. Moreover, observations will appear in practice in one large group and not in well-defined sets or sub-groups. What we shall do is to determine the basic probability and the dispersion, on the assumption that the given group of observations constitutes a Bernoullian series. We shall refer to the value of the dispersion so found as the value of the *hypothetical dispersion*, and denote it by σ_B. The value of the hypothetical dispersion can then be compared with the value of the dispersion computed directly from the observations. As an example, let us refer to the numerical example of a Lexian series considered in the preceding section, but let us suppose that we know nothing about the proportion of black and white balls in any drawing. To compute the value of the hypothetical dispersion we solve the equation $np=M$ for p, where M, the mean, is found by computation to be 5.089 and $n=10$, and obtain $p=0.5089$. Hence, the value of the hypothetical dispersion $\sigma_B=\sqrt{npq}=\sqrt{10(0.5089)(0.4911)}=1.581$. The value of the dispersion computed directly from the observations has already been found to be 2.803 ± 0.209, and the discrepancy between the two values of the dispersion indicates clearly the character of the given distribution or series. The numerical example of a Poisson series considered in the preceding section can be analyzed in the same way. It is easily verified that $p=4.580\div9=0.5089$ and that $\sigma_B=\sqrt{9(0.5089)(0.4911)}=1.500$. The value of the dispersion computed directly from the observations 1.320 ± 0.093 compared with $\sigma_B=1.500$ indicates clearly the character of that series.

We shall refer frequently to the dispersions of Bernoullian, Poisson and Lexian series—and sometimes the series themselves—as *normal*, *subnormal* and *hypernormal* respectively.

The dispersion in the first example considered above is then hypernormal and that of the second is subnormal.

It should be obvious that practically all of the problems which will arise in practice will differ from similar problems connected with games of chance, in that more factors will enter in and the basic conditions can not be controlled so easily. Statistical series will rarely conform, then, to either of the precise definitions of series given previously; the basic probability will probably be changing from observation to observation, as well as from set to set, throughout the investigation, so that the fundamental problem to be considered in practice is that of determining whether the *ultimate effect* of these changes is greater from set to set or from observation to observation within each set.

Let us ignore for the present any inequalities between what would correspond to the sizes of the various sets, and consider the number of passengers killed (m) on railroads in the United States in the decade from 1911 to 1920 inclusive.

	m	$\lvert m-M \rvert$	$(m-M)^2$
1911	299	11	121
1912	283	5	25
1913	350	38	1,444
1914	232	56	3,136
1915	199	89	7,921
1916	242	46	2,116
1917	301	13	169
1918	471	183	33,489
1919	273	15	225
1920	229	59	3,481
	2879		52,127

The total number killed (i.e., passengers, employees, etc.) varies a little from year to year in the neighborhood of 10,000. Assuming $n = 10{,}000$ we obtain as the probability that a person killed in any year will be a passenger

$$p = \frac{287.9}{10000} = 0.0288 \quad \text{and} \quad \sigma_B = \sqrt{288(0.9712)} = 16.7.$$

The dispersion computed directly from the observations is

$$\sigma=\sqrt{5212.7}=72.2\pm10.8.$$

On comparison between the two values of the dispersion, we conclude that the relative number of passengers killed from year to year is very unstable.

The values of the dispersion found above should not be regarded as reliable, because the inequality of the sizes of the sets was ignored. This fault will be considered in the next section.

EXERCISES

Ignore the possible effects of different sizes of sets of observations and investigate the normality of the following series, computing the probable error of the dispersion in each case:

1. The number of business concerns, out of 122, which reported deficits to the Federal Reserve Bank of New York were as follows:

1919	5	
1920	9	$\sigma=11.14$
1921	34	
1922	18	$\sigma_B=3.78$

2. The number of deaths from accidents in coal mines in France, omitting the results of one very disastrous catastrophe, were as follows:

1901	218	1906	163
1902	196	1907	198
1903	184	1908	171
1904	193	1909	210
1905	187	1910	194

Assume the total number of miners to be 180,000.

3. The following data represent the number of children born in Sweden, adjusted to a stationary population of 5,000,000 in accordance with a method to be explained in the next section:

1881	145,230	1891	141,070	
1882	146,630	1892	134,830	
1883	144,320	1893	136,540	
1884	149,360	1894	134,840	
1885	146,600	1895	136,820	
1886	148,270	1896	135,330	
1887	148,020	1897	132,750	
1888	143,680	1898	134,820	$M = 140, 140$
1889	138,300	1899	131,320	
1890	139,600	1900	134,460	$\sigma = 5{,}718$ $\sigma_B = 369.0$

4. The following data give the number of marriages in Denmark, adjusted to a population of 2,500,000:

1888	17,605	1897	18,676	
1889	17,622	1898	18,870	
1890	17,181	1899	18,661	
1891	17,017	1900	19,015	
1892	17,012	1901	17,870	
1893	17,676	1902	17,712	
1894	17,445	1903	17,791	$M = 18035$
1895	17,736	1904	17,895	
1896	18,239	1905	17,947	$\sigma = 588.43$ $\sigma_B = 133.81$

5. The number of still-births in Denmark (freed from certain secular fluctuations in accordance with a method discussed later) were as follows:

1888	1754	1901	1741	
1889	1826	1902	1712	
1890	1741	1903	1712	
1891	1699	1904	1718	
1892	1740	1905	1730	
1893	1726	1906	1675	
1894	1666	1907	1875	
1895	1708	1908	1765	
1896	1678	1909	1745	
1897	1784	1910	1747	$M = 1735$
1898	1779	1911	1756	
1899	1728	1912	1745	
1900	1696			$\sigma = 37.09$ $\sigma_B = 41.06$

The average annual number of births is 70,000.

6. Rough estimates of the number of deaths in New York City, adjusted to a stationary population of 2,000,000, were made as follows:

1805	52,600	1865	63,400
1815	49,400	1875	55,200
1825	51,600	1885	53,600
1835	59,600	1895	46,200
1845	60,600	1905	38,000
1855	71,200	1915	30,600

70. Basic Factors and Adjusted Series.—One of the requirements of the theory, as considered previously, which can rarely be met absolutely in practice, but which is easily met in problems connected with games of chance, is that the number of observations in each of the sets shall be equal. We ignored this requirement in the example considered in the preceding section (and in the examples given in the exercises) to simplify the explanation of the application, and also because the character of the series considered was so strongly indicated that the discrepancies in the fulfilment of this requirement would clearly have no significant effect upon the final conclusions. Let us consider the proper mode of procedure when the fulfilment of the requirement is of greater relative importance.

If we find that a certain event happens m_i times in n_i trials, we conclude that the best estimate we can make of the probability of the occurrence of the event from those observations is $p_i = m_i/n_i$. If a number of sets of trials are made and the number of trials varies from set to set, the various values of m_i may not be at all comparable. If, however, the values of the bases n_i vary only a little from each other, it may prove satisfactory to establish a common base, say n. Thus, if we multiply both the numerator and denominator of p_i by n/n_i we obtain

$$p_i = \frac{m_i \dfrac{n}{n_i}}{n_i \dfrac{n}{n_i}} = \frac{m_i \dfrac{n}{n_i}}{n}.$$

We shall call the factor n/n_i the *basic factor*; it is to be employed to *adjust* the various frequencies m_1, m_2, etc., to a common base n *when that factor differs very little from unity.* The new series so obtained will be known as the *adjusted* series. As an example, the following data give the number of marriages in thousands in the United States, the corresponding basic factors for a common base $n=730$ (thousand), and the number of divorces (in thousands) for the specified years.

Year	Marriages (thousands), n_i	Basic Factors, n/n_i	Divorces (thousands), m_i
1896	614	1.19	43
1897	622	1.17	45
1898	627	1.16	48
1899	651	1.12	51
1900	685	1.07	56
1901	717	1.02	61
1902	747	0.98	61
1903	786	0.93	65
1904	781	0.94	66
1905	805	0.91	68

The second column of the following table gives the adjusted series m'_i or the number of divorces multiplied by the corresponding basic factor.

Year	m'	$(m'-57)$		$(m'-57)^2$
1896	51	−6		36
1897	53	−4		16
1898	56	−1		1
1899	57		0	0
1900	60		3	9
1901	62		5	25
1902	55	−2		4
1903	59		2	4
1904	60		3	9
1905	61		4	16
		−13	17	120

Therefore, the mean is at

$$M=57.4,$$

$$\sigma^2=12-0.16=11.84,$$

and

$$\sigma=3.44\pm0.52.$$

The empirical probability of a divorce from a marriage is

$$p=\frac{M}{n}=\frac{57.4}{730}=0.079,$$

and

$$q=0.921.$$

Hence, the Bernoullian dispersion is

$$\sigma_B=\sqrt{npq}=\sqrt{57.4(0.921)}=7.27.$$

The given series is then clearly subnormal, and we conclude that no significant disturbing influences have affected the data from year to year.

EXERCISES

1. The number of twins was found, for each of five genealogical records, to be as follows:

	Number of Births	Twins
1.	4184	11
2.	4116	32
3.	4147	11
4.	3421	25
5.	4744	8

Adjust the series and test for normality. *Ans.* $\sigma=11\pm2.3$, $\sigma_B=4$

2. The total number of persons killed on railroads in the United States was as follows:

1911	10,396	1916	10,001
1912	10,585	1917	10,087
1913	10,964	1918	9,286
1914	10,302	1919	6,978
1915	8,621	1920	6,958

Adjust the number of passengers killed, given in the text of the preceding section, to a base of 10,000, and test the normality of the series so obtained.

3 and **4.** The following distributions give (3) the population and number of deaths due to cancer among females in the registration states of 1900, and (4) the total number of deaths occurring in New York City, and the number of deaths in hospitals in New York City. Adjust the two series and test the adjusted series for normality.

	(3)		(4)	
	Total Population	Cancer Deaths	Total Deaths	Deaths in Hospitals
1906	11,169,000	10,290	76,203	19,163
1907	11,365,000	10,870	79,205	21,444
1908	11,562,000	11,290	73,072	20,684
1909	11,758,000	11,715	74,105	21,451
1910	11,954,000	12,398	76,742	22,631
1911	12,151,000	12,634	75,423	23,466
1912	12,347,000	13,350	73,008	22,198
1913	12,544,000	13,880	73,902	22,788
1914	12,740,000	14,052	74,803	23,823
1915	12,936,000	14,472	76,193	25,095

71. Statistical Series Whose Basic Factors Would Vary too Widely from Unity.—The use of basic factors to adjust a series of frequencies for purposes of investigation, of the kind considered in the preceding section, is justified only when such factors differ little from unity. It scarcely needs to be said that statistics collected from various sources are only too likely to be based upon individual investigations which will differ quite widely from each other in the number of cases investigated. Suppose, for example, that we were investigating the trend of mortality due to a certain disease throughout many countries from year to year, and we wished particularly to determine whether the mortality statistics from the various countries were comparable or not; it is obvious that the basic factors of two countries, one of which has a population, say, one hundred times that of the other, would differ too widely, and that the use of such factors would be equivalent to weight-

ing the statistics of the smaller country one hundred times as much as the statistics of the larger country. As a matter of fact, the mean error of the adjusted statistics of the smaller country would probably be ten times that of the adjusted statistics of the larger country. (Explain.) On the other hand, if we should weight the terms of such a statistical series according to the relative values of their precision (i.e., proportional to the square root of their bases), we should thereby assign a character of stability to the statistics with the larger bases which would probably be unjustified.

The method which we shall employ will be simply stated without any demonstrations.[3] It is based upon the two relations

$$\sigma_B = \sqrt{\frac{Nn}{\Sigma n_i}}\sqrt{npq} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (59)$$

and

$$\sigma = \sqrt{\frac{\pi}{2}}\,\text{A. D.} \quad \left(\sqrt{\frac{\pi}{2}} = 1.2533\right), \quad . \quad . \quad . \quad (60)$$

of which the first was derived for the specific purpose of analyzing the types of distributions under consideration, the second is a relation between the dispersion and average deviation given in a previous exercise (Exercise 8, p. 178), and both relations are supposed to hold only *when the given distribution is Bernoullian.* Since, however, the average deviation gives no special emphasis to extreme deviations, the variations in its value in distributions which are not Bernoullian should not only be in the direction in which we should expect the dispersion to vary, but should reflect to a minimum extent the inconsistencies introduced in comparing frequencies with widely different bases. We shall therefore employ the second relation to estimate the dispersion which would be computed directly from the statistics if the number of observations in each set

[3] Fisher, p. 159.

were the same. It is well to note in that connection that the average deviation of such a series would be

$$\text{A. D.} = \frac{\Sigma n_i \left| \frac{n}{n_i} m_i - np \right|}{\Sigma n_i} = \frac{n\Sigma \mid m_i - n_i p \mid}{\Sigma n_i},$$

and that for either relation

$$p = \frac{\Sigma m_i}{\Sigma n_i}.$$

Relation (59) is then to be employed to determine what would correspond to the hypothetical dispersion.

It should be noted that the application of the two relations given above does not require that the terms of a distribution or series be adjusted. As an example, the number of automobiles and the number of automobile fatalities in ten states in 1920 were as follows:

	Number of Automobiles,	Fatalities		
	n_i	m_i	$n_i p$	$\mid m_i - n_i p \mid$
California	681,000	734	593	141
Illinois	663,000	728	577	151
Indiana	400,000	248	348	100
Kentucky	127,000	84	110	25
Michigan	476,000	419	415	4
Minnesota	325,000	178	283	105
Missouri	346,000	231	301	70
Nebraska	239,000	104	208	104
Ohio	723,000	717	630	87
Tennessee	117,000	130	102	28
	$\Sigma n_i = 4{,}097{,}000$	$\Sigma m_i = 3574$		815

Hence,

$$p = \frac{3574}{4{,}097{,}000} = 0.0009,$$

$$\text{A. D.} = \frac{1000 \times 815}{4{,}097{,}000} = 0.1989,$$

and

$$\sigma = 1.2533 \times \text{A. D.} = 0.249 \pm 0.037.$$

Also

$$\sigma_B^2 = \frac{10 \times 1000}{4{,}097{,}000} \times 1000(0.0009)(0.9991),$$

or

$$\sigma_B = 0.047.$$

We therefore conclude that the ratio of fatalities to the number of automobiles varies considerably from state to state. As a matter of fact, the divergence between the values of the dispersions would probably have been much greater if we had included the data for states such as New York, New Jersey, etc., where a large part of the automobile traffic is urban.

EXERCISES

Test the normality of the following series:

1. The following data give the number of wives and the number of childless wives for different periods, according to statistics collected from 22 genealogical records of American families:

	Total	Childless
1750–99	1966	37
1800–49	5530	225
1850–69	3062	181
1870–79	1086	88

2. The following data give the total number of accidents resulting in permanent disability and the total number of fatal accidents, for various countries.

		Permanent Disability	Fatal
Austria	(1897–1906)	82,446	8,349
Belgium	(1905–1908)	8,204	1,838
Denmark	(1899–1906)	4,192	389
France	(1899–1908)	140,877	18,708
Germany	(1899–1908)	313,219	59,893
Italy	(1898–1902)	9,701	2,224
Norway	(1895–1905)	4,496	832
Russia	(1904–1906)	34,981	2,345
		598,116	84,578

72. The Lexian Ratio and the Charlier Coefficient.—It has been found from experience, and it would be only natural to suspect that a great majority of statistical series are hypernormal or Lexian, or that the ultimate effect of the change of the basic probability during an investigation is less than the ultimate effect of the change of the probability from investigation to investigation. This fact constitutes the main reason why we should never be too hasty in placing absolute faith in the value of an important probability which has been determined empirically. Any plan, then, for testing the reliability and comparability of a set of ratios or probabilities, like the one which we have been considering in this chapter, deserves considerable attention and possible extension. Results obtained by the plan considered here may be expressed better and more concisely by two formulas. The simpler one is the ratio

$$L=\frac{\sigma}{\sigma_B}, \quad \ldots\ldots\ldots \quad (61)$$

which is called the *Lexian ratio*. It is evident that a series is normal, hypernormal or subnormal according as $L=1$, $L>1$ or $L<1$ respectively.

As the expression

$$\frac{\Sigma(p_i-p)^2}{N},$$

marks the essential difference between the dispersions of a normal and a hypernormal series, it seems only natural to employ it as a measure of the variations in the chances from the mean p. Since, however, it is dependent on the absolute values of these chances, we divide it by the square of the mean of these chances. Denoting the quotient by ρ^2 we have

$$\rho^2=\frac{\Sigma(p_i-p)^2}{Np^2}.$$

We next replace the numerator by its value taken from the formula for the dispersion of a Lexian series where, however,

we replace σ_L by σ, since the latter is to be computed directly from the given observations. We have then

$$\rho^2 = \frac{\sigma^2 - \sigma_B^2}{(n^2 - n)p^2}.$$

Neglecting n in comparison with n^2, and remembering that $M_B = np$, we have as an approximation

$$\rho = \frac{\sqrt{\sigma^2 - \sigma_B^2}}{M_B}. \quad . \quad . \quad . \quad . \quad . \quad (62)$$

100ρ is called the *Charlier coefficient of disturbancy* of a series. It is evident that the Charlier coefficient is zero for a normal series, imaginary for a subnormal series and real for a hypernormal series.

It is easily verified that the Lexian ratio for the series of divorces considered previously is

$$L = \frac{3.44}{7.27} = 0.473$$

and that the Charlier coefficient is

$$100\rho = 100\,\frac{\sqrt{11.84 - 52.89}}{57.4},$$

or imaginary.

EXERCISES

1. Compute the Charlier coefficient for the following distributions, giving the number of deaths from accidents in coal mines in various countries:

	England	Germany	United States	Belgium	Austria	France	Japan
1901	1224	1170	1982	164	81	218	263
1902	1116	995	2263	150	73	196	188
1903	1134	960	1952	160	50	184	278
1904	1116	900	2135	130	62	193	239
1905	1215	930	2214	127	99	187	354
1906	1161	985	2944	133	70	1262	578
1907	1179	1240	2977	144	73	198	399
1908	1188	1355	2220	150	58	171	262
1909	1287	1021	2440	133	73	210	667
1910	1530	985	2391	133	63	194	245

The total number of miners in each of the countries during the designated period was approximately: England, 900,000; Germany, 500,000; United States, 610,000; Belgium, 140,000; Austria, 68,000; France, 180,000; Japan, 110,000.

Answers:

England, 4.71; Germany, 9.27; United States, 12.07; Belgium, 2.55; Austria, 13.85; France, 34.19; Japan, 44.12.

2. The Courrieres mine disaster, in 1906, accounted for 1099 of the deaths tabulated above, for France. Eliminate the results of this disaster and show that the remaining statistical series is subnormal.

73. Series with Certain Secular Fluctuations.—So far we have given our attention solely to the detection of disturbing influences in a statistical series. There may, however, be certain well-known influences in a given series, whose effects are already fairly well appreciated, which we should like to eliminate sufficiently to permit us to investigate the presence of other but disturbing and possibly vitiating influences. Thus, the investigation of a series giving the number of deaths by months in a community, due to a disease which occurs more frequently at certain times of the year than at others, would, of course, verify the presence of violently disturbing influences. The fluctuations in such a case would be periodic and might well conceal for the time the presence of other and undesirable influences. To take another illustration, it is fairly well established that the death rate due to cancer is on the increase. Hence, a series giving the number of such deaths in a community by successive years would, of course, prove to be hypernormal if the increase in death rate proved to be significant, and again the fluctuations due to this increase might well conceal the presence of other and undesirable influences. It will be necessary to omit consideration of methods of eliminating periodic fluctuations, since such consideration would necessitate the inclusion of methods of determining the period of such fluctuations. We shall give our attention, then, to the second type of fluctuations and shall

refer to them as *secular* fluctuations. In fact, we shall limit our discussion to secular fluctuations due to influences which tend to work uniformly in the same direction—that is, to give either a constant increase or a constant decrease in the series of frequencies to be considered. We propose first, then, to derive a method of measuring the rate of such a secular fluctuation, and then to show how these fluctuations may be eliminated sufficiently to allow an investigation of the presence of other and disturbing influences. It is evident that a series showing, say, an apparent increase in the death rate due to a certain disease may, after the secular fluctuations due to this increase in death rate are eliminated, prove so unreliable as to practically vitiate any satisfactory conclusions that might otherwise be drawn from the series taken as a whole.

We shall assume that the fundamental probability varies by a constant difference k from one set of observations to another, so that

$$p_i = p_{i-1} + k,$$

and

$$p_i = p_1 + (i-1)k.$$

It is easily verified that the average p of N such probabilities is

$$p = p_1 + \frac{N-1}{2}k,$$

and that

$$p_i - p = \left(i - \frac{N+1}{2}\right)k.$$

Now, if we assume the observed terms $m_1, m_2, \ldots m_N$ of a given statistical series to be essentially the same as the values which would be given by the corresponding probabilities $p_1, p_2, \ldots p_N$, or the same as $np_1, np_2, \ldots np_N$ where we assume the number of observations necessary to determine any observed term to be always the same or n, the relation just obtained may be written

$$m_i - M = \left(i - \frac{N+1}{2}\right)nk, \quad . \quad . \quad . \quad . \quad . \quad (A)$$

where M denotes the mean or average of the observed terms.

The average variation or fluctuation from one value of m_i to the next is evidently

$$nk = \frac{1}{N-1}\{(m_2-m_1)+(m_3-m_2)+\ldots+(m_N-m_{N-1})\}$$

$$= \frac{m_N-m_1}{N-1}. \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (B)$$

Thus, given a column of values of $i(=1, 2, \ldots N)$ and a column of corresponding values of m_i, we need only to subtract the first value of m_i from the last value and divide the difference by the total number of values diminished by unity to obtain nk.

To remove the secular fluctuations we need, then, to set up a new column of values of m_i obtained by solving relation (A) for M or

$$M = m_i - \left(i - \frac{N+1}{2}\right)nk, \quad \ldots\ldots \quad (62)$$

and letting $i=1, 2, \ldots N$ successively. We shall refer to this final series as the *residual* series.

If the number of observations is not the same for all sets, the series will have to be adjusted in accordance with the schemes suggested previously. As an illustration, let us consider the following distribution, which gives the number (m) of deaths from cancer in the City of New York adjusted to a basic population of $n=1{,}000{,}000$. The last column gives the residual series.

Year	i	m	m'
1904	1	609	664
1905	2	639	681
1906	3	619	649
1907	4	658	676
1908	5	631	637
1909	6	683	677
1910	7	710	698
1911	8	710	686
1912	9	721	685
1913	10	718	670

It is easily verified that the adjusted distribution or series is hypernormal.

By formula (B):

$$nk=\frac{718-609}{9}=12.1.$$

If we substitute this value in formula (62), and let $i=1, 2, \ldots 10$ successively, we obtain the residual series given to the right above.

It is now easily verified that the new series is subnormal, or that the Lexian ratio is

$$L=\tfrac{17}{26} \text{ (approximately)},$$

and that the Charlier coefficient is imaginary.

EXERCISES

Remove the secular fluctations in the following statistical series and then test the residual series for normality:

1. The following data give the number of still-births in Denmark corresponding to an assumed stationary total number of births of 70,000:

Year		Year		Year		Year		Year	
1888	1861	1893	1788	1898	1797	1903	1685	1908	1694
1889	1924	1894	1719	1899	1737	1904	1682	1909	1665
1890	1830	1895	1753	1900	1696	1905	1705	1910	1658
1891	1779	1896	1714	1901	1732	1906	1602	1911	1658
1892	1811	1897	1811	1902	1694	1907	1723	1912	1638

Ans. $nk=-8.92$; $\sigma=37.09$; $\sigma_B=41.6$; and the Charlier coefficient is imaginary.

2. The following table gives the number of deaths from accidents in coal mines, in the United States, in which less than five men were killed. Assume the total number of miners to be 630,000:

1900	1843	1905	1964	1910	2085
1901	1863	1906	2075	1911	1984
1902	1837	1907	2190	1912	1839
1903	1768	1908	1967	1913	1957
1904	1911	1909	2053	1914	1810

Ans. C. C. $=5.51$

3. The following data give the number of deaths from cancer in New York City, as reduced to a stationary population of 1,000,000:

1889	377	1894	423	1899	513	1904	609	1909	683
1890	476	1895	442	1900	547	1905	639	1910	710
1891	410	1896	493	1901	595	1906	619	1911	710
1892	444	1897	505	1902	540	1907	658	1912	721
1893	462	1898	515	1903	580	1908	631	1913	718

Ans. Residual series is slightly subnormal.

4. The following data give the number of deaths among members of the Brotherhood of Locomotive Firemen and Engineers:

	Members	Deaths
1904	54,434	453
1905	55,287	496
1906	58,849	461
1907	62,916	581
1908	66,408	436
1909	65,315	411
1910	73,469	519
1911	79,942	522
1912	85,292	558

CHAPTER X

CORRELATION THEORY

74. Introduction.—Suppose that the average person were asked whether sons of fathers who lived to advanced ages also tended, in the long run, to live to advanced ages, etc. The ordinary procedure would be to try to recall actual examples. If the experience consisted merely of a few pairs of fathers and sons both of whom, in each case, lived to advanced ages, the answer would probably be in the affirmative. If the experience consisted of a few cases showing the opposite tendency, the answer would probably be in the negative. In neither case, however, would the answer be conclusive, because the experience would be entirely too limited. Correlation theory is highly useful in such a problem, because it enables one to assimilate and weigh any amount of experience, however large, in a single application, and hence to give what is generally accepted as a fairly conclusive answer. A measure of this correlation will often serve to establish either a connection between two phenomena which had been suggested only by statistical data and whose nature might still be unknown, or the independence of two phenomena which had been regarded hitherto as related in some way. Correlation theory has proved useful in almost every conceivable field, including biology, psychology, education, etc., but particularly in problems of heredity, and has helped to indicate what characters are inherited and what characters are due to peculiar environment.

Correlation theory is very useful in suggesting causal relationship between characters. Thus, if a high death rate [1]

[1] The student should be cautioned from the outset about using *rates*

due to one disease in a community is almost invariably accompanied by a high death rate due to a second disease, a causal relationship is suggested, furnishing a problem for medical authorities whose solution might lead to the discovery of causes hitherto unsuspected. It should be emphasized, however, that the responsibility for the final explanation in such an investigation rests not with the statistician but with the authority versed in the particular field.

Two characters are said to be correlated when with a selected value of one, certain values of the other are likely to be associated. Stated more precisely, *two characters are said to be correlated if to a selected series of values of the one there correspond values of the other whose mean values are functions of the selected values.* The full meaning of this statement will be fully appreciated and understood when we have entered more fully into the theory.

75. The Correlation Surface.—Let the probability of the occurrence of a deviation or error x occurring be

$$\frac{1}{\sigma_x\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma_x^2}},$$

and the probability of the occurrence of a deviation y be the same expression with y substituted for x, where x and y are used also as subscripts of the corresponding standard deviations for purposes of distinction. If, and only if, the two probabilities are independent, the probability of their joint occurrence is

$$z_i=\frac{1}{\sigma_x\sigma_y\cdot 2\pi}e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)}. \quad . \quad . \quad . \quad . \quad (63)$$

which may be regarded as the equation of a surface. It is easily verified that all sections of this surface parallel to the xz-plane would be normal curves having the same standard

(generally known as "indices") themselves in correlation problems, lest his results prove to be *spurious*. See Art. 81.

deviation (σ_x) and all sections parallel to the yz-plane would be normal and would have the same standard deviation (σ_y).

It should be evident that the equation of the surface representing the probability of the joint occurrence of the deviations x and y, when the latter are not necessarily independent, would have to be of a more general form than that given above and one which would have to reduce to that given above as a special case. It has been shown elsewhere [2] that certain very reasonable assumptions lead to the form

$$z = ke^{-\frac{1}{2}(ax^2+2hxy+by^2)}, \quad . \quad . \quad . \quad . \quad . \quad . \quad (64)$$

wherein the essential difference from the preceding form is the presence of the xy-term in the exponent. The complete presentation of the derivation of equation (64) would carry us too far from our present purpose. However, if we seek merely a more general form of (63) it seems only natural to assume form (64). Why is it unnecessary to include $x-$, $y-$ and constant terms in the exponent of (64)? We shall now proceed to express the coefficients a, b, h and k in terms of characters which can be readily computed.

Let us recall for a moment certain features connected with the determinations of areas under curves and volumes under surfaces by the calculus. In determining the area under a curve by integration, the integrand represents the *typical* ordinate or strip of area of infinitesimal width whose area is summed. Likewise, if we seek to determine the volume under a surface by double integration, the integrand of the first integration represents as before the *typical* ordinate and the integrand of the second integration represents the *typical* section of infinitesimal thickness of the surface. Hence, if we integrate (64) with respect to x from $-\infty$ to $+\infty$ we obtain the *typical* y section or distribution. Performing this integration, we obtain

[2] See Elderton's "Frequency Curves and Correlation," p. 109.

$$k\int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2+2hxy+by^2)}dx = ke^{-\frac{a}{2}\left(by^2-\frac{h^2y^2}{a^2}\right)}\int_{-\infty}^{\infty} e^{-\frac{a}{2}\left(x^2+\frac{2h}{a}xy+\frac{h^2y^2}{a^2}\right)}dx$$

(see equation (34)) $$= k\sqrt{\frac{2\pi}{a}}e^{-\frac{y^2}{2}\left(b-\frac{h^2}{a}\right)} = k\sqrt{\frac{2\pi}{a}}e^{-\frac{y^2}{2\sigma_y^2}},$$

and for this typical section or distribution

$$\frac{1}{\sigma_y^2} = b - \frac{h^2}{a} = b\left(1-\frac{h^2}{ab}\right).$$

Similarly, integrating the same expression with respect to y, we obtain

$$\frac{1}{\sigma_x^2} = a\left(1-\frac{h^2}{ab}\right).$$

If we let

$$r = -\frac{h}{\sqrt{ab}}, \quad \ldots\ldots\ldots \quad (A)$$

we obtain

$$\frac{1}{\sigma_x^2} = a(1-r^2),$$

or

$$a = \frac{1}{\sigma_x^2(1-r^2)}. \quad \ldots\ldots\ldots \quad (B)$$

Similarly

$$b = \frac{1}{\sigma_y^2(1-r^2)}. \quad \ldots\ldots\ldots \quad (C)$$

Hence

$$h = -r\sqrt{ab} = \frac{-r}{\sigma_x\sigma_y(1-r^2)}. \quad \ldots\ldots \quad (D)$$

If we complete the double integration begun above (say, with respect to x) we obtain the volume or N, or

$$N = k\sqrt{\frac{2\pi}{a}}\int_{-\infty}^{\infty} e^{-\frac{y^2}{2\sigma_y^2}}dy = k\sqrt{\frac{2\pi}{a}}\cdot\sigma_y\sqrt{2\pi}$$

$$= \frac{2\pi k\sigma_y}{\sqrt{a}} = 2\pi k\sigma_x\sigma_y\sqrt{1-r^2},$$

or

$$k = \frac{N}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}}. \quad \ldots\ldots\ldots \quad (E)$$

Substituting the values of a, h, b and k just found in (64), the equation becomes

$$z=\frac{N}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}}e^{-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_x^2}-\frac{2xyr}{\sigma_x\sigma_y}+\frac{y^2}{\sigma_y^2}\right)}. \quad . \quad . \quad (65)$$

The expression just found for z with $N=1$ represents the probability of the joint occurrence of the deviations x and y whether they be dependent or independent. The expression for z_i represents the probability of the joint occurrence of x and y only when they are independent. The difference between the two expressions depends solely upon r; in fact, for $r=0$ the expression for z reduces to that for z_i. All this suggests the advisability of using values of r as measures of the dependence or correlation of the two characters. The surface corresponding to equation (65) is called the *correlation surface* and r is called the *correlation coefficient*. In the next section we shall derive a method for computing the value of the correlation coefficient.

76. The Product Moment and the Formula for the Correlation Coefficient.—Just as the ordinate at the mean or the centroid vertical of a plane curve is the ordinate about which the first moment is zero, the centroid vertical of a surface is the vertical line about which the first moments of both the x and y distributions are zero (i.e., the vertical line passes through the center of gravity).

If we define what is generally called the *product moment* of a surface $z=f(x, y)$ by the relation

$$\Sigma xy=\int\int xyf(x, y)dxdy,$$

valued between appropriate limits, then the product moment of the correlation surface about the centroid vertical is

$$\Sigma xy=k\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xye^{-\frac{1}{2}(ax^2+2hxy+by^2)}dxdy$$

$$=k\int_{-\infty}^{\infty}ydy\int_{-\infty}^{\infty}xe^{-\frac{1}{2}(ax^2+2hxy+by^2)}dx.$$

But the integration with respect to x may be written

$$-\frac{1}{a}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2+2hxy+by^2)}(-ax-hy+hy)dx$$

$$=-\frac{1}{a}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2+2hxy+by^2)}(-ax-hy)dx-\frac{hy}{a}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2+2hxy+by^2)}dx.$$

It is left for the student to show that the value of the first integral is zero. Moreover, the value of the second integral has already been found (verify) to be $\sqrt{\frac{2\pi}{a}}e^{-\frac{y^2}{2\sigma_y^2}}$. Hence, the value of the product moment of the correlation surface reduces to

$$\Sigma xy=k\int_{-\infty}^{\infty} y\left(-\frac{hy}{a}\right)\left(\sqrt{\frac{2\pi}{a}}e^{-\frac{y^2}{2\sigma_y^2}}\right)dy$$

$$=-\frac{hk\sqrt{2\pi}}{a\sqrt{a}}\int_{-\infty}^{\infty} y^2e^{-\frac{y^2}{2\sigma_y^2}}\,dy.$$

It is easily verified that

$$\int_{-\infty}^{\infty} y^2e^{-\frac{y^2}{2\sigma_y^2}}\,dy=\sigma_y^3\sqrt{2\pi}.$$

(Integrate by parts, letting $u=y$)
Hence,

$$\Sigma xy=\frac{-hk\sqrt{2\pi}\sigma_y^3\sqrt{2\pi}}{a\sqrt{a}}.$$

If we substitute the values of a, h and k in this expression, we obtain

$$\Sigma xy=N\sigma_x\sigma_y r,$$

whence

$$r=\frac{\Sigma xy}{N\sigma_x\sigma_y}. \qquad \ldots\ldots\ldots \quad (66)$$

Explain how Σxy may mean the same as $\Sigma xyf(x, y)$; compare the distinction between the two with the distinction between the ordinary arithmetic and the weighted arithmetic averages.

77. The Unit Product Moment. Notation.—At this point it is well to recall the distinction made previously between a frequency surface and the corresponding histogram, and to keep in mind that the various moments of a histogram are only approximately equal to the corresponding moments of the corresponding curve or surface, the various moments of the histogram being found usually by ordinary summation or simple addition, and the moments of the curve or surface being usually found by means of the calculus. Thus, $\Sigma xyf(x, y)$ (or Σxy) for a histogram means that each value of $f(x, y)$ is to be multiplied by the corresponding values of x and y (class marks) and the sum of all such products found. For example, in the correlation table given in Art. 79, $f(x, y)$ for $x = -1$ and $y = -2$ has the value 11 and for that frequency $xyf(x, y) = (-1)(-2)11 = 22$. The sum of all such products is called the product moment of the histogram. However, the value of such a product moment would not be the value of the product moment about the centroid vertical. Just as it was found very inconvenient in frequency distributions of two dimensions to compute the values of the moments directly from deviations from the mean, and more convenient to employ a trial mean and make corrections later, so in the case of the product moment it will prove more convenient to employ a trial centroid vertical and employ corrections to give the value of the product moment about the true centroid vertical. The formula for making these corrections will be derived in the next section.

If we designate the *unit* product moment about the centroid vertical of a *surface* by μ_{xy} (and about any other vertical line by μ'_{xy}) formula (66) becomes

$$r = \frac{\mu_{xy}}{\sigma_x \sigma_y} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (67)$$

The unit product moment about the centroid vertical of a *histogram* is denoted by ν_{xy} (and about any other vertical line by ν'_{xy}).

78. The Standard Method of Computing the Product Moment about the Centroid Vertical.—We shall now establish the relation between the product moment of a surface about the centroid vertical and the product moment about any other vertical line.

Regarding this case as analogous to that of two dimensions, let $z_r = f_r(x, y)$ be any value of $z = f(x, y)$, x and y the corresponding coordinates with respect to the centroid vertical, and x' and y' the corresponding coordinates with respect to any other vertical line taken as the z-axis. Let the component distances of the centroid vertical from the arbitrary z-axis be h and k. Then, we have

$$x' = x + h \quad \text{or} \quad x = x' - h,$$

$$y' = y + k \quad \text{or} \quad y = y' - k.$$

Hence, if we write $f(x, y)$ after the bracket for sake of brevity—where, however, it is to be included in the summation

$$\mu_{xy} = \frac{\Sigma xyf(x, y)}{N} = \frac{1}{N}(\Sigma x'y' - k\Sigma x' - h\Sigma y' + hk\Sigma)f(x, y)$$

$$= \frac{1}{N}(\Sigma x'y' - k\Sigma(x + h) - h\Sigma(y + k) + hk\Sigma)f(x, y)$$

$$= \frac{1}{N}(\Sigma x'y' - k\Sigma x - h\Sigma y - hk\Sigma)f(x, y).$$

But $\Sigma xf(x, y) = \Sigma yf(x, y) = 0$. (Why?)

Also $\Sigma f(x, y) = N$. (Why?) and $hk\Sigma f(x, y) = hkN$.

Therefore,

$$\mu_{xy}=\frac{\Sigma x'y'}{N}-hk=\mu'_{xy}-hk, \quad . \quad . \quad . \quad . \quad (68)$$

where μ'_{xy} is the unit product moment of a surface about any vertical line and h and k are, for purposes of computation, *the first unit moments of the* x *and* y *distributions respectively.* Strictly speaking, the symbol μ would be appropriate only when the summation denoted by Σ is performed by integration. However, as quadrature formulas are almost never used in computing the value of the correlation coefficient, a confusion of symbols in such a connection is not serious. Formula (67) can now be written in the more workable form

$$r=\frac{\mu'_{xy}-hk}{\sigma_x\sigma_y}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (69)$$

79. The Computation of the Correlation Coefficient.—We shall now show by illustration the computation of the correlation coefficient where the only feature which is essentially new is the computation of the product moment. The numbers given in the table itself record the number of students in the freshman class of a certain college having the corresponding weights and abdominal measurements designated. Thus, the "13" in the third row from the bottom refers to 13 students who weighed approximately 110 pounds and had abdominal measurements of approximately 24 inches. It is desired to compute the value of the correlation coefficient for these two characteristics.

First, the rows are summed horizontally to give the y-distribution and the columns vertically to give the x-distribution. Then the various moments of these distributions are computed exactly in accordance with methods previously considered. The product moment is computed in each of two ways for purposes of check—by rows and by columns. For

example, the number 268 (in the last column) $= -2\{13(-3)+42(-2)+11(-1)+1(0)\}$. The entire computation is given.[3]

$$h = \frac{-412}{455} = -0.905,$$

$$\sigma'_x = \sqrt{\frac{966}{455} - (-0.905)^2} = 1.142,$$

$$k = \frac{-169}{455} = -0.363,$$

$$\sigma'_y = \sqrt{\frac{793}{455} - (-0.363)^2} = 1.201,$$

Wts. (lbs.)	Abdominal Measurements (inches) 24	26	28	30	32	34	36	38	40	z	y	yz	y^2z	xy
230								2		2	6	12	72	48
215									1	1	5	5	25	25
200					1				1	2	4	8	32	24
185				1	1	1	1	1		5	3	15	45	30
170			1	12	5	3	2			23	2	46	92	32
155			11	29	13	1				54	1	54	54	4
140		10	85	38	2					135	0			
125	1	62	88	11						162	−1	−162	162	215
110	13	42	11	1						67	−2	−134	268	268
95	1	2								3	−3	− 9	27	21
80		1								1	−4	− 4	16	8
z	15	117	196	92	22	5	3	3	2	455		−169	793	675
x	− 3	−2	−1	0	1	2	3	4	5					
xz	−45	234	196		22	10	9	12	10	−412				
											$N = 455$			
x^2z	135	468	196		22	20	27	48	50	966				
xy	90	312	97		30	20	21	60	45	675				

[3] Crelle's (multiplication) tables and Barlow's tables of squares, square roots, reciprocals, etc., will be found very useful in computing the value of a correlation coefficient.

$$\nu'_{xy}=\frac{675}{455}=1.484,$$

$$\nu_{xy}=1.484-(-0.905)(-0.363)=1.153,$$

$$r=\frac{1.153}{(1.142)(1.201)}=0.843.$$

The formula for the probable error of the correlation coefficient is

$$\frac{\pm 0.6745(1-r^2)}{\sqrt{N}}, \quad . \quad . \quad . \quad . \quad . \quad . \quad (70)$$

which, in this case, has the value ± 0.022.

The means of the two distributions are

$$M_x= 30+ 2(-0.905)= 28.190,$$

$$M_y=140+15(-0.363)=134.76.$$

The true values of the standard deviation should also take consideration of the unit of measurement. The true values would then be

$$\sigma_x=2(1.142)=2.284 \quad \text{and} \quad \sigma_y=15(1.201)=18.015.$$

80. The Units of Measurement.—It will surely be noticed that the original units of measurement were completely ignored in the computation of the value of the correlation coefficient in the preceding section; the units of measurement, which were 2 and 15, were changed at once to unity, and the origin of each distribution was translated to the class mark nearest the mean, as determined by inspection. It is permissible, in computing the value of the correlation coefficient, to ignore the original units of measurement, because if they were retained they would cancel in the final result; for, suppose that the unit of measurement in the x direction were m and in the y direction n; then h, obtained by assuming the unit of measurement to be unity, would actually be mh, and ν'_2 would be $m^2\nu'_2$; therefore ν_2 would be $m^2\nu'_2-m^2h^2$ and σ_x would be $m\sigma_x$. Likewise σ_y would be $n\sigma_y$. The product moment μ'_{xy} would be

$mn\mu'_{xy}$ and $\mu_{xy}=\mu'_{xy}-hk$ would be $mn\mu'_{xy}-mnhk=mn\mu_{xy}$. Therefore

$$r=\frac{mn\mu_{xy}}{mn\sigma_x\sigma_y}=\frac{\mu_{xy}}{\sigma_x\sigma_y},$$

which shows that the value of the correlation coefficient is independent of the units of measurement m and n. It should be emphasized, however, that this statement holds merely for the correlation coefficient; in computing the standard deviation or in locating the mean of a distribution, the unit of measurement is essential. Thus, the mean of the abdominal measurements is $30-2(0.905)=28.190$ and the standard deviation is $2(1.142)=2.284$.

It will be shown later that the correlation coefficient may have any value between -1 and 1. Zero signifies no correlation, and 1 perfect positive correlation; that is, when $r=1$, deviations of either characteristic are definitely associated with like deviations—in size and sign—of the other. Perfect negative correlation signifies a definite association of deviations but of opposite sense. It should perhaps be added that the best way for the student to become familiar with the kinds of data which are appropriate for a problem in correlation and with the method of arranging such data in a table would be to undertake an *original* problem.

EXERCISES

Find the value of the correlation coefficient and its probable error for the following tables:

1. Heights in inches and weights in pounds of Glasgow school-boys, ages 4.5 to 5.5 years. (From Biometrika, Vol. XI.)

	Weight					
Height	26	31	36	41	46	51
31	2					
34	5	15	5			
37	1	18	72	8		
40		5	87	90	7	1
43			4	35	21	5
46			1		2	

2. Use the correlation table of grades in mathematics and by psychological test given in Art. 40.

Ans. $M_x = 149.98$ $\sigma_x = 19.8$
$M_y = 74.28$ $\sigma_y = 10.8$ $r = 0.37 \pm 0.03$

3. Heights and weights of Glasgow school-boys, ages 13.5 to 14.5 years.

Height	Weight															
	46	51	56	61	66	71	76	81	86	91	96	101	106	111	116	121
43		1														
46		2	3	1												
49	1	1	5	7	7	1										
52		2	3	16	31	36	9	1								
55				7	27	41	49	27	6	2	2					
58			1			7	19	28	28	16	7	3				
61							1	1	5	4	4	5	5			
64									1		2	1	1			1

4. Left cubits (mm.) and left middle finger (mm.) Cairo-born Egyptians. (Biom. Vol. XII.)

mm.	Cubits										
	395	410	425	440	455	470	485	500	515	530	545
94.5			1								
98.5	2	1	1	1	1						
102.5	1	9	17	4	1						
106.5		6	33	59	13						
110.5		4	18	62	75	11	6				
114.5			1	29	100	68	17				
118.5			1	5	30	75	21	5			
122.5					4	27	36	11	2		
126.5						5	19	8	4		
130.5							1	2	2		
134.5									2		1

5. Relationship between size of annual income and per cent of total annual expenditure for food—from an investigation into the standard of living in the District of Columbia.

	Per Cent									
	22	26	30	34	38	42	46	50	54	58
$600			1		2	3	4	4	1	1
800	1	1		5	5	9	12	10	5	1
1000			3	2	11	6	4	5		
1200		1	2	5	8	13	2	2	1	
1400	2		4	8	6	8	6	1		
1600	2	3	4	8	3	2	3			
1800		1		2		1	1	1		
2000		3								
2200			1							

81. Regression. Linear Regression.—We shall refer to the columns (or rows) of frequencies of a correlation table corresponding to a particular value of x as a y-array and to the rows (or columns) of frequencies corresponding to a particular value of y as x-arrays. Then, if we plot the means of the various y-arrays of an extensive table, these means will lie approximately on a smooth curve. The means can then be fitted by a curve by any convenient method such as that of least squares. The curve thus obtained is called the *curve of regression* (from the original application which showed the tendency of individual traits to "regress" or conform back to those of the general population), and if this curve is a straight line the *regression* is said to be *linear*. As the regression is usually linear we shall consider this case only.

The immediate problem is to determine the values of m and b for which the line $y=mx+b$ will fit the means of the y-arrays. The problem can be solved for each particular correlation table, but it is possible and preferable to express m and b once for all in terms of familiar moments which can be readily computed or read off from computations already made in any particular problem. If we employ the method of least squares we are to determine the values of m and b for which

$$\Sigma(\bar{y}-mx-b)^2$$

is a minimum, where $\bar{y}$ is the mean of the y's of any y-array corresponding to the class mark x, and x and y are deviations from the means of the corresponding distributions. Differentiating with respect to b and m, we have respectively

$$-2\Sigma\ (\bar{y}-mx-b)=0, \quad \text{or} \quad \Sigma\bar{y}=m\Sigma x+\Sigma b. \quad . \quad (F)$$

$$-2\Sigma x(\bar{y}-mx-b)=0, \quad \text{or} \quad \Sigma x\bar{y}=m\Sigma x^2+b\Sigma x. \quad . \quad (G)$$

If we take the origin at the centroid

$$\Sigma\bar{y}=\Sigma x=0. \quad \text{(Why?)}$$

Therefore, by (F)

$$\Sigma b=Nb=0, \quad \text{or} \quad b=0,$$

and by (G)

$$\Sigma x\bar{y}=m\Sigma x^2,$$

or

$$\frac{\Sigma x\bar{y}}{N}=m\frac{\Sigma x^2}{N}=m\sigma_x^2. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (H)$$

Now, *for any* y-*array*, x is constant and $\Sigma x\bar{y}$ can be written $x\Sigma\bar{y}$, which has the same value as $x\Sigma y$ (Why?) or Σxy, and since the values of $\Sigma x\bar{y}$ and Σxy are the same for each y-array they are the same for the entire correlation table. Hence, $\Sigma x\bar{y}$ or Σxy is the product moment and (H) may be written

$$\nu_{xy}=m\sigma_x^2,$$

whence

$$m=\frac{\nu_{xy}}{\sigma_x^2}=r\frac{\sigma_y}{\sigma_x},$$

and the equation of the line of regression may be written

$$y=r\frac{\sigma_y}{\sigma_x}x. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (71)$$

Similarly, the line of regression which fits the means of the x-arrays is

$$x=r\frac{\sigma_x}{\sigma_y}y,$$

which, of course, is not in general equivalent to the equation (71).

The expressions $r\frac{\sigma_y}{\sigma_x}$ and $r\frac{\sigma_x}{\sigma_y}$ are called *regression coefficients.* It is very important to note that although the value of the correlation coefficient is independent of the original units of measurement, the values of the standard deviation and, hence, of the regression coefficients are not. The value of the regression coefficient computed, assuming the units of measurement to be unity instead of, say, m and n, is easily adjusted by multiplying by n/m (or m/n).

82. Interpretation of the Regression and Correlation Coefficients.—The regression coefficient is highly useful not only for purposes of statistical measurement but also as providing the best interpretation of the correlation coefficient. If we regard, for the moment, the arithmetic average of a set of variates as the most probable value of the variates, a selected value of x need only be multiplied by the value of the regression coefficient to determine the most probable value of y to be associated with it. As the value of the regression coefficient $r\frac{\sigma_y}{\sigma_x}$, in the problem considered in a preceding section has the value $\frac{15\times 1.201}{2\times 1.142}(0.843)=6.649$, the most probable value of y to be associated with a selected value of x is $6.649x$. Thus, an abdominal measurement of 34 inches corresponds to a deviation from the mean of such measurements of $34.000-28.190$ or 5.810. Hence, the most probable deviation from the mean of the weights to be associated with the deviation 5.810 is $6.649(5.810)$ or 38.63 which corresponds to a weight of $134.76+38.63=173.39$ pounds.

If the equation of the line of regression be written

$$\frac{y}{\sigma_y}=r\frac{x}{\sigma_x},$$

it is evident that if we think of the deviations x and y as

expressed in standard units by dividing by the value of the standard deviation, the regression coefficient becomes identically the correlation coefficient and the interpretation of the regression coefficient given above applies to the correlation coefficient.

EXERCISES

1. Find the psychological grade which corresponds to a mathematical grade of 70. *Ans.* 147.

2. Find the mathematical grade which corresponds to a psychological grade of 135.

3. Find the abdominal measurement which corresponds to a weight of 180 pounds.

4. Find the weight which corresponds to an abdominal measurement of 40 inches.

83. Upper and Lower Bounds of the Correlation Coefficient. —The sum of the squares of the residuals found in a correlation table by taking the difference between the observed deviation y and the corresponding theoretical value as given by the equation of the line of regression (71) is

$$\Sigma\left(y-r\frac{\sigma_y}{\sigma_x}x\right)^2=\Sigma y^2-2r\frac{\sigma_y}{\sigma_x}\Sigma xy+r^2\frac{\sigma_y^2}{\sigma_x^2}\Sigma x^2$$

$$=N\sigma_y^2-2r^2\frac{\sigma_y}{\sigma_x}N\sigma_x\sigma_y+r^2\frac{\sigma_y^2}{\sigma_x^2}N\sigma_x^2$$

$$=N\sigma_y^2(1-r^2),$$

and since the left side must be positive, so must the right side, which requires that $r^2\leq 1$ or that r be not greater than $+1$ or less than -1. If $r^2=1$ all the points corresponding to the frequencies lie on the regression lines and *the lines coincide.* (Why?) In other words, if the value of the deviation of one character is given when $r=\pm 1$, the value of the corresponding deviation of the other character is uniquely determined.

84.[4] **Correlation between n Variates.**—We have already discussed correlation between two characters or variates and explained the method of computing the correlation coefficent and of obtaining the equation of the line of regression. It is sometimes desirable to measure the correlation between several variates. Thus, it may be illuminating to measure the correlation between the index prices of foods with other factors of our social life, or between characteristics not only of a father and his son but also of the son and both parents, and even of the brothers and sisters and of grandparents, and to determine a linear relation corresponding to the line of regression in the case of two variates. The regression coefficients would at least indicate the relative influence of the several factors. The formulas necessary for treating n variates, which correspond to the formulas derived previously, will now be given but without derivation. The application of the formulas will involve no new principles except possibly a few elementary ones involving determinants.

The normal correlation function of which (65) is a special case where $n=2$ has the form

$$z=\frac{1}{\sigma_1\sigma_2 \ldots \sigma_n\sqrt{S(2\pi)^n}}\, e^{-\frac{1}{2S}\sum\frac{S_{ij}}{\sigma_i\sigma_j}x_ix_j} \quad . \quad . \quad . \quad . \quad (72)$$

where S is the determinant

$$\begin{vmatrix} 1 & r_{12} & r_{13} \ldots r_{1n} \\ r_{21} & 1 & r_{23} \ldots r_{2n} \\ r_{31} & r_{32} & 1 \ldots r_{3n} \\ \cdot & \cdot & \cdot \ldots \cdot \\ r_{n1} & r_{n2} & r_{n3} \quad 1 \end{vmatrix} \quad . \quad . \quad . \quad . \quad . \quad (73)$$

and S_{ij} is the co-factor of r_{ij} $(=r_{ji})$ or the determinant obtained

[4] This article may well be omitted in an elementary course. Some general explanation would, however, be desirable to explain the use of formula (75) to avoid spurious results.

by striking out the row and column passing through r_{ij} and giving it the proper sign. The various correlation coefficients in (73) are computed in the ordinary way but are called *partial* correlation coefficients because in computing their values the other variates are regarded for the time as constant.

In order to consider cases where n is greater than 2, we introduce the symbols

$$\alpha_i = \frac{\sigma_i}{\sqrt{S_{ii}}}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (A)$$

$$R_{ij} = -\frac{S_{ij}}{\sqrt{S_{ii}S_{ij}}}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (74)$$

R_{ij} is called the *multiple* correlation coefficient between the characters (i.e., deviations) x_i and x_j of the nth order where there are n characters involved. It is easily verified that R_{ij} is identically the correlation coefficient r when $n=2$. Likewise, when $n=3$,

$$R_{12} = \frac{r_{12} - r_{23}r_{31}}{\sqrt{(1-r_{23}^2)(1-r_{31}^2)}}. \quad . \quad . \quad . \quad . \quad (75)$$

Referring again to the case of n characters and following the same line of procedure to obtain formula (71), the following regression equation of the first degree, for determining the most probable value of x_1 to be associated with selected values of $x_2, x_3, \ldots x_n$, is obtained:

$$x_1 = R_{12}\frac{\alpha_1}{\alpha_2}x_2 + R_{13}\frac{\alpha_1}{\alpha_3}x_3 + \ldots + R_{1n}\frac{\alpha_1}{\alpha_n}x_n, \quad . \quad . \quad (76)$$

which may be written

$$x_1 = b_{12}x_2 + b_{13}x_3 + \ldots + b_{1n}x_n, \quad . \quad . \quad . \quad . \quad (77)$$

where

$$b_{1i} = R_{1i}\frac{\alpha_1}{\alpha_i} = -\frac{\sigma_1}{\sigma_i}\cdot\frac{S_{1i}}{S_{11}}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (78)$$

It should be recalled that in computing the values of the partial correlation coefficients r_{1i} the original units of measure-

ment may be ignored, but that they must be retained in expressing the values of the standard deviations as they appear in formula (78).

The labor involved in computing the multiple correlation coefficients proves excessive if expansions of formula (74) are used. Formulas,[5] called *recursion formulas*, have been derived for expressing the value of a multiple correlation coefficient in terms of those of the next lower order but the easiest method, when the order is greater than 3, is to valuate the various co-factors of determinant (73) by successive expansion in accordance with Laplace's development, after zeros have been made to appear in the first column (or row). The method is given in any textbook on elementary determinant theory and will not be reproduced here. If Crelle's (multiplication) tables are used, a determinant of the sixth order involving partial correlation coefficients to three decimals can, with a little practice, be valuated in about twenty minutes, and determinants of lower order in less time.

As a numerical illustration, the values of the standard deviations and correlation coefficients between the mean temperatures of the months of (1) July, (2) June, . . . (5) March, computed for Lund, Sweden, by Charlier, are as follows:

$\sigma_1 = 1.91$

$\sigma_2 = 1.99 \quad r_{12} = 0.734$

$\sigma_3 = 1.99 \quad r_{13} = 0.518 \quad r_{23} = 0.586$

$\sigma_4 = 1.80 \quad r_{14} = 0.361 \quad r_{24} = 0.409 \quad r_{34} = 0.429$

$\sigma_5 = 2.68 \quad r_{15} = 0.146 \quad r_{25} = 0.209 \quad r_{35} = 0.303 \quad r_{45} = 0.421$

[5] Yule's recursion formula is

$$R_{ab(abc\ldots hk)} = \frac{R_{ab} - R_{ak}R_{bk}}{\sqrt{(1-R_{ak}^2)(1-R_{bk}^2)}}$$

where $R_{ab} = R_{ab(abc\ldots h)}$

$R_{ak} = R_{ak(ac\ldots hk)}$

$R_{bk} = R_{bk(bc\ldots hk)}$

Substituting the values of these correlation coefficients in the determinant (73), the values of the co-factors S_{1i} $(i=1, 2, \ldots 5)$ were found according to the method outlined above to be

$$\begin{array}{llll} S_{11}= & 0.410 & & \\ S_{12}=-0.264 & \text{whence} & b_{12}= & 0.617 \\ S_{13}=-0.053 & & b_{13}= & 0.125 \\ S_{14}=-0.027 & & b_{14}= & 0.070 \\ S_{15}= & 0.024 & b_{15}=-0.041 & \end{array}$$

Since

$$S=S_{11}+r_{12}S_{12} \ldots +r_{15}S_{15},$$

or

$$1=\frac{\Sigma r_{1i}S_{1i}}{S},$$

the value of the expression on the right was computed, as a check, to be 1.0006 which is satisfactory, considering that the values of the correlation coefficients were computed to only three decimals.

The regression equation (77) giving the most probable value of the mean temperature in the month of (1) July may now be written

$$x_1=0.617x_2+0.125x_3+0.070x_4-0.041x_5,$$

where x_2 represents the average temperature in the month of June of a selected year diminished by the mean temperature obtained from a large number of years for this month; that is, x_2 is a selected deviation from the mean. The quantities x_3, x_4, and x_5 have analogous meanings relating to the months of May, April and March, respectively. The quantity x_1 represents the most probable deviation of the temperature of July from the mean temperature of that month. Similar regression equations could be obtained from the original determinant S for the other months.

It should be added that for a *complete* solution of the problem of general forecasting of temperature all other meteorological factors should be included in the investigation.

85. Spurious Correlation.—There is one type of correlation problem whose characteristics should be carefully noted and which can be treated in only one way to avoid "spurious" results. It is the type whose data consist of "indices" or whose pairs of measurements have been obtained by dividing by a common divisor. The frequencies of such a correlation table may show correlation which is due partly or wholly to the effect of the common divisor. A complete treatment of the type can not be given here and is unnecessary, for a single example can be given which will suffice to justify our contention and which will illustrate clearly the possible effect of the presence of such common divisors. If the following hypothetical data were arranged in the usual form of a correlation table:

x	y	$f(x, y)$	x	y	$f(x, y)$
40	30	8	50	40	2
20	30	2	40	50	2
30	30	4	30	20	2
50	30	4	40	20	4
60	30	2	50	20	2
30	40	2	40	10	2
40	40	4			

the table would appear as follows:

		x				
		20	30	40	50	60
y	50			2		
	40		2	4	2	
	30	2	4	8	4	2
	20		2	4	2	
	10			2		

It is evident from the appearance of the table that the data were selected with the direct purpose of showing no correlation;

for, the values of the product moment and, therefore, of the correlation coefficient would be zero. But suppose that each pair of measurements were divided by some number, say, according to the following plan:

Divisors	x	y	$f(x, y)$	New x	New y
10	40	30	8	4	3
5	20	30	2	4	6
6	30	30	4	5	5
10	50	30	4	5	3
6	60	30	2	10	5
10	30	40	2	3	4
20	40	40	4	2	2
5	50	40	2	10	8
10	40	50	2	4	5
5	30	20	2	6	4
20	40	20	4	2	1
5	50	20	2	10	4
10	40	10	2	4	1
			40		

the new measurements and frequencies, when arranged in a correlation table, would appear as follows:

y \ x	2	3	4	5	6	7	8	9	10
8									2
7									
6			2						
5			2	4					2
4		2			2				2
3			8	4					
2	4								
1	4		2						

It is obvious, without any attempt at computation, that there is now considerable correlation. It is evident, however, that the correlation is due entirely to the use of the common divisors.

As a concrete illustration of a problem which would probably lead to spurious correlation, suppose that we were investigating statistical evidences for the relationship between the effects of cancer and of diabetes, as indicated by the comparison of the death rates due to these diseases in a large number of communities. We wish to know whether, in the long run, the size of the death rate due to one disease is associated with particular sizes of death rates due to the other disease. Suppose that we ascertain the pair of death rates for each of a large number of communities and find that the number of communities showing a particular pair of death rates appears as a frequency in the correlation table and the deviations from the mean of the two death rates or their class marks as values of x and y. It is obvious that to determine the two death rates for a community the number of deaths for both diseases must be divided by the population of the community, and that the value of the correlation coefficient is very apt to prove spurious.

Spurious results in such problems may be avoided by treating the divisors as values of a third variate and using formula (75) for multiple correlation. Thus, in the problem concerning the correlation between cancer and diabetes it is necessary to determine the values of the three correlation coefficients between the three characters, "deaths (not death rates) due to cancer," "deaths due to diabetes" and "population" taken in pairs. The values of these three coefficients should then be substituted in formula (75) to give the desired value of the (multiple) correlation coefficient.

86. One Criticism of the Use of the Correlation Coefficient. Suggestions for Further Study.—Students who are sufficiently interested in the analysis of statistics to continue the study further should be informed of one defect in the use of the correlation coefficient which may prove important in some

investigations. The means of a set of arrays may lie approximately on a curve which is not well covered by the preceding discussion, and there may be much greater correlation in such a case than the value of the correlation coefficient would indicate; in fact, it is easy to formulate hypothetical situations in which the correlation is perfect but the value of the correlation coefficient differs quite significantly from unity. Such situations occur very rarely in practice, and it was deemed inadvisable to include a treatment of them in this volume. It may be sufficient to say that what is called the correlation *ratio* has been found to give better measures of the correlation under such circumstances.

The student is advised to extend his study also to the derivations of the various formulas for probable errors which he will find in Biometrika, the Drapers Research Memoirs and other journals, besides the textbooks on statistics by Yule, Bowley, Fisher, Elderton, Jones, etc. Considerable information is contained in the preface of Pearson's "Tables for Statisticians."

The systematic fitting of frequency curves together with suitable quadrature formulas is very important, particularly the remarkable system of curves known as the Pearson frequency curves, a brief account of which may be found either in Elderton's "Frequency Curves and Correlation," or Jones' "First Course in Statistics," as well as in Pearson's "Tables." The work of statisticians of northern Europe should not be omitted in that connection, but unfortunately a reading knowledge of the languages of that part of Europe is necessary for the complete study of that work. Arne Fisher's "Mathematical Theory of Probabilities" constitutes the best and almost the only account of that work in English and includes also a treatment of many other interesting and fundamental problems, particularly those connected directly with probabilities. A brief account is given there also of a subject in the field of statistics which is little appreciated as yet in this country—that of invariants.

LOGARITHMS

N	0	1	2	3	4	5	6	7	8	9
10	000000	004321	008600	012837	017033	021189	025306	029384	033424	037427
11	041393	045323	049218	053078	056905	060698	064458	068186	071882	075547
12	079181	082785	086360	089905	093422	096910	100371	103804	107210	110590
13	113943	117271	120574	123852	127105	130334	133539	136721	139879	143015
14	146128	149219	152288	155336	158362	161368	164353	167317	170262	173186
15	176091	178977	181844	184691	187521	190332	193125	195900	198657	201397
16	204120	206826	209515	212188	214844	217484	220108	222716	225309	227887
17	230449	232996	235528	238046	240549	243038	245513	247973	250420	252853
18	255273	257679	260071	262451	264818	267172	269513	271842	274158	276462
19	278754	281033	283301	285557	287802	290035	292256	294466	296665	298853
20	301030	303196	305351	307496	309630	311754	313867	315970	318063	320146
21	322219	324282	326336	328380	330414	332438	334454	336460	338456	340444
22	342423	344392	346353	348305	350248	352183	354108	356026	357935	359835
23	361728	363612	365488	367356	369216	371068	372912	374748	376577	378398
24	380211	382017	383815	385606	387390	389166	390935	392697	394452	396199
25	397940	399674	401401	403121	404834	406540	408240	409933	411620	413300
26	414973	416641	418301	419956	421604	423246	424882	426511	428135	429752
27	431364	432969	434569	436163	437751	439333	440909	442480	444045	445604
28	447158	448706	450249	451786	453318	454845	456366	457882	459392	460898
29	462398	463893	465383	466868	468347	469822	471292	472756	474216	475671
30	477121	478567	480007	481443	482874	484300	485721	487138	488551	489958
31	491362	492760	494155	495544	496930	498311	499687	501059	502427	503791
32	505150	506505	507856	509203	510545	511883	513218	514548	515874	517196
33	518514	519828	521138	522444	523746	525045	526339	527630	528917	530200
34	531479	532754	534026	535294	536558	537819	539076	540329	541579	542825
35	544068	545307	546543	547775	549003	550228	551450	552668	553883	555094
36	556303	557507	558709	559907	561101	562293	563481	564666	565848	567026
37	568202	569374	570543	571709	572872	574031	575188	576341	577492	578639
38	579784	580925	582063	583199	584331	585461	586587	587711	588832	589950
39	591065	592177	593286	594393	595496	596597	597695	598791	599883	600973
40	602060	603144	604226	605305	606381	607455	608526	609594	610660	611723
41	612784	613842	614897	615950	617000	618048	619093	620136	621176	622214
42	623249	624282	625312	626340	627366	628389	629410	630428	631444	632457
43	633468	634477	635484	636488	637490	638489	639486	640481	641474	642465
44	643453	644439	645422	646404	647383	648360	649335	650308	651278	652246
45	653213	654177	655138	656098	657056	658011	658965	659916	660865	661813
46	662758	663701	664642	665581	666518	667453	668386	669317	670246	671173
47	672098	673021	673942	674861	675778	676694	677607	678518	679428	680336
48	681241	682145	683047	683947	684845	685742	686636	687529	688420	689309
49	690196	691081	691965	692847	693727	694605	695482	696356	697229	698101

N	0	1	2	3	4	5	6	7	8	9
50	698970	699838	700704	701568	702431	703291	704151	705008	705864	706718
51	707570	708421	709270	710117	710963	711807	712650	713491	714330	715167
52	716003	716838	717671	718502	719331	720159	720986	721811	722634	723456
53	724276	725095	725912	726727	727541	728354	729165	729974	730782	731589
54	732394	733197	733999	734800	735599	736397	737193	737987	738781	739572
55	740363	741152	741939	742725	743510	744293	745075	745855	746634	747412
56	748188	748963	749736	750508	751279	752048	752816	753583	754348	755112
57	755875	756636	757396	758155	758912	759668	760422	761176	761928	762679
58	763428	764176	764923	765669	766413	767156	767898	768638	769377	770115
59	770852	771587	772322	773055	773786	774517	775246	775974	776701	777427
60	778151	778874	779596	780317	781037	781755	782473	783189	783904	784617
61	785330	786041	786751	787460	788168	788875	789581	790285	790988	791691
62	792392	793092	793790	794488	795185	795880	796574	797268	797960	798651
63	799341	800029	800717	801404	802089	802774	803457	804139	804821	805501
64	806180	806858	807535	808211	808886	809560	810233	810904	811575	812245
65	812913	813581	814248	814913	815578	816241	816904	817565	818226	818885
66	819544	820201	820858	821514	822168	822822	823474	824126	824776	825426
67	826075	826723	827369	828015	828660	829304	829947	830589	831230	831870
68	832509	833147	833784	834421	835056	835691	836324	836957	837588	838219
69	838849	839478	840106	840733	841359	841985	842609	843233	843855	844477
70	845098	845718	846337	846955	847573	848189	848805	849419	850033	850646
71	851258	851870	852480	853090	853698	854306	854913	855519	856124	856729
72	857333	857935	858537	859138	859739	860338	860937	861534	862131	862728
73	863323	863917	864511	865104	865696	866287	866878	867467	868056	868644
74	869232	869818	870404	870989	871573	872156	872739	873321	873902	874482
75	875061	875640	876218	876795	877371	877947	878522	879096	879669	880242
76	880814	881385	881955	882525	883093	883661	884229	884795	885361	885926
77	886491	887054	887617	888179	888741	889302	889862	890421	890980	891537
78	892095	892651	893207	893762	894316	894870	895423	895975	896526	897077
79	897627	898176	898725	899273	899821	900367	900913	901458	902003	902547
80	903090	903633	904174	904716	905256	905796	906335	906874	907411	907949
81	908485	909021	909556	910091	910624	911158	911690	912222	912753	913284
82	913814	914343	914872	915400	915927	916454	916980	917506	918030	918555
83	919078	919601	920123	920645	921166	921686	922206	922725	923244	923762
84	924279	924796	925312	925828	926342	926857	927370	927883	928396	928908
85	929419	929930	930440	930949	931458	931966	932474	932981	933487	933993
86	934498	935003	935507	936011	936514	937016	937518	938019	938520	939020
87	939519	940018	940516	941014	941511	942008	942504	943000	943495	943989
88	944483	944976	945469	945961	946452	946943	947434	947924	948413	948902
89	949390	949878	950365	950851	951338	951823	952308	952792	953276	953760
90	954243	954725	955207	955688	956168	956649	957128	957607	958086	958564
91	959041	959518	959995	960471	960946	961421	961895	962369	962843	963316
92	963788	964260	964731	965202	965672	966142	966611	967080	967548	968016
93	968483	968950	969416	969882	970347	970812	971276	971740	972203	972666
94	973128	973590	974051	974512	974972	975432	975891	976350	976808	977266
95	977724	978181	978637	979093	979548	980003	980458	980912	981366	981819
96	982271	982723	983175	983626	984077	984527	984977	985426	985875	986324
97	986772	987219	987666	988113	988559	989005	989450	989895	990339	990783
98	991226	991669	992111	992554	992995	993436	993877	994317	994757	995196
99	995635	996074	996512	996949	997386	997823	998259	998695	999131	999565

ANTILOGARITHMS

L	0	1	2	3	4	5	6	7	8	9
.00	100000	100231	100462	100693	100925	101158	101391	101625	101859	102094
.01	102329	102565	102802	103039	103276	103514	103753	103992	104232	104472
.02	104713	104954	105196	105439	105682	105925	106170	106414	106660	106905
.03	107152	107399	107647	107895	108143	108393	108643	108893	109144	109396
.04	109648	109901	110154	110408	110662	110917	111173	111429	111686	111944
.05	112202	112460	112720	112980	113240	113501	113763	114025	114288	114551
.06	114815	115080	115345	115611	115878	116145	116413	116681	116950	117220
.07	117490	117761	118032	118304	118577	118850	119124	119399	119674	119950
.08	120226	120504	120781	121060	121339	121619	121899	122180	122462	122744
.09	123027	123310	123595	123880	124165	124451	124738	125026	125314	125603
.10	125893	126183	126474	126765	127057	127350	127644	127938	128233	128529
.11	128825	129122	129420	129718	130017	130317	130617	130918	131220	131522
.12	131826	132130	132434	132739	133045	133352	133660	133968	134276	134586
.13	134896	135207	135519	135831	136144	136458	136773	137088	137404	137721
.14	138038	138357	138676	138995	139316	139637	139959	140281	140605	140929
.15	141254	141579	141906	142233	142561	142889	143219	143549	143880	144212
.16	144544	144877	145211	145546	145881	146218	146555	146893	147231	147571
.17	147911	148252	148594	148936	149279	149624	149968	150314	150661	151008
.18	151356	151705	152055	152405	152757	153109	153462	153815	154170	154525
.19	154882	155239	155597	155955	156315	156675	157036	157398	157761	158125
.20	158489	158855	159221	159588	159956	160325	160694	161065	161436	161808
.21	162181	162555	162930	163305	163682	164059	164437	164816	165196	165577
.22	165959	166341	166725	167109	167494	167880	168267	168655	169044	169434
.23	169824	170216	170608	171002	171396	171791	172187	172584	172982	173380
.24	173780	174181	174582	174985	175388	175792	176198	176604	177011	177419
.25	177828	178238	178649	179061	179473	179887	180302	180717	181134	181552
.26	181970	182390	182810	183231	183654	184077	184502	184927	185353	185780
.27	186209	186638	187068	187499	187932	188365	188799	189234	189671	190108
.28	190546	190985	191426	191867	192309	192752	193197	193642	194089	194536
.29	194984	195434	195884	196336	196789	197242	197697	198153	198609	199067
.30	199526	199986	200447	200909	201372	201837	202302	202768	203236	203704
.31	204174	204644	205116	205589	206063	206538	207014	207491	207970	208449
.32	208930	209411	209894	210378	210863	211349	211836	212324	212814	213304
.33	213796	214289	214783	215278	215774	216272	216770	217270	217771	218273
.34	218776	219280	219786	220293	220800	221309	221820	222331	222844	223357
.35	223872	224388	224905	225424	225944	226464	226986	227510	228034	228560
.36	229087	229615	230144	230675	231206	231739	232274	232809	233346	233884
.37	234423	234963	235505	236048	236592	237137	237684	238232	238781	239332
.38	239883	240436	240991	241546	242103	242661	243220	243781	244343	244906
.39	245471	246037	246604	247172	247742	248313	248886	249459	250035	250611
.40	251189	251768	252348	252930	253513	254097	254683	255270	255859	256448
.41	257040	257632	258226	258821	259418	260016	260615	261216	261818	262422
.42	263027	263633	264241	264850	265461	266073	266686	267301	267917	268534
.43	269153	269774	270396	271019	271644	272270	272898	273527	274157	274789
.44	275423	276058	276694	277332	277971	278612	279254	279898	280543	281190
.45	281838	282488	283139	283792	284446	285102	285759	286418	287078	287740
.46	288403	289068	289734	290402	291072	291743	292415	293089	293765	294442
.47	295121	295801	296483	297167	297852	298538	299226	299916	300608	301301
.48	301995	302691	303389	304089	304789	305492	306196	306902	307610	308319
.49	309030	309742	310456	311172	311889	312608	313329	314051	314775	315500

L	0	1	2	3	4	5	6	7	8	9
.50	316228	316957	317687	318420	319154	319890	320627	321366	322107	322849
.51	323594	324340	325087	325837	326588	327341	328095	328852	329610	330370
.52	331131	331894	332660	333426	334195	334965	335738	336512	337287	338065
.53	338844	339625	340408	341193	341979	342768	343558	344350	345144	345939
.54	346737	347536	348337	349140	349945	350752	351560	352371	353183	353997
.55	354813	355631	356451	357273	358096	358922	359749	360579	361410	362243
.56	363078	363915	364754	365595	366438	367282	368129	368978	369828	370681
.57	371535	372392	373250	374111	374973	375837	376704	377572	378443	379315
.58	380189	381066	381944	382825	383707	384592	385478	386367	387258	388150
.59	389045	389942	390841	391742	392645	393550	394457	395367	396278	397192
.60	398107	399025	399945	400867	401791	402717	403645	404576	405509	406443
.61	407380	408319	409261	410204	411150	412098	413048	414000	414954	415911
.62	416869	417830	418794	419759	420727	421697	422669	423643	424620	425598
.63	426580	427563	428549	429536	430527	431519	432514	433511	434510	435512
.64	436516	437522	438531	439542	440555	441570	442588	443609	444631	445656
.65	446684	447713	448745	449780	450817	451856	452898	453942	454988	456037
.66	457088	458142	459198	460257	461318	462381	463447	464515	465586	466659
.67	467735	468813	469894	470977	472063	473151	474242	475335	476431	477529
.68	478630	479733	480839	481948	483059	484172	485289	486407	487528	488652
.69	489779	490908	492040	493174	494311	495450	496592	497737	498884	500035
.70	501187	502343	503501	504661	505825	506991	508159	509331	510505	511682
.71	512861	514044	515229	516416	517607	518800	519996	521195	522396	523600
.72	524807	526017	527230	528445	529663	530884	532108	533335	534564	535797
.73	537032	538270	539511	540754	542001	543250	544503	545758	547016	548277
.74	549541	550808	552077	553350	554626	555904	557186	558470	559758	561048
.75	562341	563638	564937	566239	567545	568853	570164	571479	572796	574116
.76	575440	576766	578096	579429	580764	582103	583445	584790	586138	587489
.77	588844	590201	591562	592925	594292	595662	597035	598412	599791	601174
.78	602560	603949	605341	606736	608135	609537	610942	612350	613762	615177
.79	616595	618016	619441	620869	622300	623735	625173	626614	628058	629506
.80	630957	632412	633870	635331	636796	638263	639735	641210	642688	644169
.81	645654	647143	648634	650130	651628	653131	654636	656145	657658	659174
.82	660693	662217	663743	665273	666807	668344	669885	671429	672977	674528
.83	676083	677642	679204	680769	682339	683912	685488	687068	688652	690240
.84	691831	693426	695024	696627	698232	699842	701455	703072	704693	706318
.85	707946	709578	711214	712853	714496	716143	717794	719449	721107	722770
.86	724436	726106	727780	729458	731139	732825	734514	736207	737904	739605
.87	741310	743019	744732	746449	748170	749894	751623	753356	755092	756833
.88	758578	760326	762079	763836	765597	767361	769130	770903	772681	774462
.89	776247	778037	779830	781628	783430	785236	787046	788860	790679	792501
.90	794328	796159	797995	799834	801678	803526	805378	807235	809096	810961
.91	812831	814704	816582	818465	820352	822243	824138	826038	827942	829851
.92	831764	833681	835603	837529	839460	841395	843335	845279	847227	849180
.93	851138	853100	855067	857038	859014	860994	862979	864968	866962	868960
.94	870964	872971	874984	877001	879023	881049	883080	885116	887156	889201
.95	891251	893305	895365	897429	899498	901571	903649	905733	907821	909913
.96	912011	914113	916220	918333	920450	922571	924698	926830	928966	931108
.97	933254	935406	937562	939723	941890	944061	946237	948418	950605	952796
.98	954993	957194	959401	961612	963829	966051	968278	970510	972747	974990
.99	977237	979490	981748	984011	986279	988553	990832	993116	995405	997700

INDEX

(Numbers refer to pages)

A

Adjusted (statistical) series, 194
Algebraic treatment of symbols, 29
Arrays, 221
Artillery, field, 157
Asymmetrical curves, 167
Average, arithmetic (or mean), 79
  weighted, 80
  deviation, 97
  geometric, 81
    weighted, 81
  harmonic (mean), 81
  median (quartiles, percentiles, etc.), 81
  mode, 85

B

Basic factors, 194
Bernoulli numbers, 31
  series, 178
Beta function, 60
Bias of index numbers, 107

C

Centroid, 119
  vertical, 119
Charlier check, 110
  coefficient of disturbancy, 202
Class marks, 83
  limits, 83
  interval, 84
Coefficient, correlation, 212
    partial, 226
    multiple, 226
Coefficient, of disturbancy, 202
  regression, 223
  of variation, 99
Combinations, 65
Computation, numerical, 5
Consistency of observations, 95
Correlation (theory), 208
  coefficient, 212
    computation, 216
    probable error of, 218
  spurious, 209, 229
  between n variates, 225
  lagging, 132
  ratio, 232
  surface, 209, 212
Curve, fitting, 85
  by moments, 112
  by least squares, 149
Cycles (time series), 132
Cyclical fluctuations (time series), 131

D

Determinants, 225, 227
Deviation, 88
  average, 97
  standard, 94
Difference, finite, 13
  leading, 15
Dispersion, 94, 123
  hypothetical, 190
Distributions, frequency, 83, 88
  normal, 93, 137

(Numbers refer to pages)

E

Elimination of secular trend, 127
of seasonal variation, 130
Errors, absolute and relative, 2
compensating and accumulative, 3
extended meaning of, 90
maximum, 156
probable, 153
maximum, 156
standard, 156
Expectation, mathematical, 76
of life, 54

F

Factorial, 17–18
Field artillery, 157
Frequency curves, 85
Pearson, 139
distributions, 83, 88
surfaces, 88
Functions, Beta, 30
Gamma, 56
rational, 27
rational integral, 17

G

Gamma function, 56
Graduations of frequencies, 85, 113, 114, 126, 141–145
Guiding principle of probable error, 161

H

Histogram, rectangular, 86, 88
Homogeneity of populations, 70
Hypothetical dispersion, 190

I

Indeterminate forms, 57
Index numbers, 105
Indices in correlation, 209, 229
Integration, finite, 21
by parts, 26
by substitution, 57
Interpolation by Newton's formula, 34
by Lagrange's formula, 36
by leading-difference formulas, 40, 44, 47
tangential, 42
of ordinates among arcas, 45
of areas, 48
Interpretation of correlation coefficient, 224
of regression coefficient, 223

L

Lag of time series, 131
Lagrange's interpolation formula, 36
Least squares, 148
curve fitting, 149
Lexian series, 178
ratio, 201

M

Maximum error, 156
Mean or arithmetic average, 79
harmonic, 81
provisional or trial, 88
Moments, definition, 109
curve fitting, 112
unit, 117
about mean, 119
product, 125, 212
unit, 214
of point binomial, 173
Mortality tables, abridged, 51
Multiple correlation, 226

N

Newton's formula, 19
Normal curve, 136

(Numbers refer to pages)

Normal curve, derivation of equation, 139
graduations, 141–145
tables of ordinates and areas, 145
(subnormal, hypernormal) series, 190

P

Partial correlation coefficient, 226
differentiation, 149
Point binomial, 167
application of, 173
Poisson series, 178
Probability, *a priori*, 64
empirical or *a posteriori*, 70
factors, 157
Probable error in a single observation, 153, 155
maximum, 156
of various quantities, 160
Problems, some famous, 76

Q

Quadrature formulas, 118

R

Reason, cogent, 74
insufficient, 74
Recursion formulas, 227
Regression, 221, 226
coefficient, 223
equation, 221
Relatives, price, 103
link, 103
chain, 104
Residual (statistical) series, 205
Root-mean-square, 94, 123

S

Seasonal variation, 127
Secular trend of a time series, 124
fluctuations of a statistical series, 203
Series, time, 102
statistical (Bernoullian, Poisson, Lexian), 178
normal, etc., 190
adjusted, 194
residual, 205
Skewness of frequency curves, 167
Standard deviation, 94, 123
error, 156
Statistical series, 178
Summation of series, 22, 32

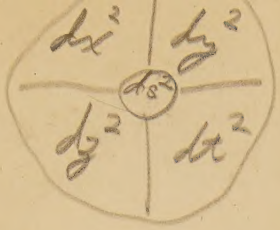
T

Taylor's expansion, 30
Time series, 102
cyclical fluctuations, 131
secular trend, 124
seasonal variation, 27

V

Variability, 99

1.	.5	.32	.14	.04
3.5	1.75	1.12	.49	.14
8.	4.00	2.56	1.12	.32
12.5	6.25	4.00	1.75	.5
	12.5	8.	3.5	1.



Please return this book to

for

He who takes what isn't his
be apt to find himself in jeopardy

and

She who takes what isn't
Beware
May dodge the process
but in the end it
burns